

Metadata and Ubiquitous Access to Culture, Science and Digital Humanities • DC

1-4 Sep 2015 • São Paulo, Brazil

**2015 Proceedings of the International
Conference on Dublin Core and Metadata
Applications**

Proceedings Edited by:

Mariana Curado Malta

CEISE/ISCAP - Polytechnic of Oporto, Portugal & Algoritmi

Research Center - University of Minho, Portugal

Silvana A. Borsetti Gregorio Vidotti

Universidade Estadual Paulista (UNESP), Brazil

Published by:

Dublin Core Metadata Initiative

a project of ASIS&T

Conference Host:



**São Paulo, Brazil, USA
1-4 September 2015**

ISSN: 1939-1366 (Online)



WORKSHOPS

- DC-1**, Dublin, Ohio USA — 1-3 March 1995
- DC-2**, Warwick, U. K— 1-3 April 1996
- DC-3**, Dublin, Ohio USA — 24-25 September 1996
- DC-4**, Canberra, Australia — 3-5 March 1997
- DC-5**, Helsinki, Finland — 6-8 October 1997
- DC-6**, Washington D.C. USA — 2-4 November 1998
- DC-7**, Frankfurt, Germany — 25-27 October 1999
- DC-8**, Ottawa, Canada — 4-6 October 2000

CONFERENCES

- DC-2001**, Tokyo, Japan — 22-26 October 2001
- DC-2002**, Florence, Italy — 14-17 October 2002
- DC-2003**, Seattle, Washington, USA — 28 September - 2 October 2003
- DC-2004**, Shanghai, China — 10-14 October 2004
- DC-2005**, Leganés (Madrid), Spain — 12-15 September 2005
- DC-2006**, Manzanillo, Colima, Mexico — 3-6 October 2006
- DC-2007**, Singapore — 27-31 August 2007
- DC-2008**, Berlin, Germany — 22-26 September 2008
- DC-2009**, Seoul, Korea — 12-16 October 2009
- DC-2010**, Pittsburgh, Pennsylvania, USA — 20-22 October 2010
- DC-2011**, The Hague, The Netherlands — 21-23 September 2011
- DC-2012**, Kuching, Sarawak, Malaysia — 3-7 September 2012
- DC-2013**, Lisbon, Portugal — 2-6 September 2013
- DC-2014**, Austin, Texas — 8-11 October 2014
- DC-2015**, São Paulo, Brazil — 1-4 September 2015

© **DCMI 2015**

Copyright for individual articles is retained by the authors with first publication rights granted to DCMI for publication in print and electronic proceedings. By virtue of their appearance in this open access publication, articles are free to be used with proper attribution of the author for educational and other non-commercial purposes. Other uses may require the permission of the authors.

ISSN: 1939-1366 (Online)

São Paulo, Brazil



DC-2015 Welcome

Welcome to DC 2015 in São Paulo, Brazil! On behalf of the Dublin Core Metadata Initiative (DCMI), it is my pleasure to join our host, the Universidade Estadual Paulista (UNESP), and the Organizing Team in welcoming one and all to the Dublin Core DC-2015 Conference and Annual Meeting.

The annual *International Conference on Dublin Core and Metadata Applications* has established itself as one world's premiere gathering of researchers, practitioners, implementers, and students active in cultural sector and learning-related metadata. This gathering marks the 15th International Conference series which was launched with the initial conference in Tokyo in October 2001. This is also the first time the event has been held in South America, and only the second time in a Portuguese-speaking country (DC-2013 was held in Lisbon).

The theme of this year's conference, *Metadata and Ubiquitous Access to Culture, Science and Digital Humanities*, speaks to the need for structured metadata to support ubiquitous access across the Web to the treasure troves of resources spanning cultures, in science, and in the digital humanities. That need is now common knowledge among information systems designers and implementers. To this end, DC-2015 will feature tutorials, special sessions, papers, project reports, and posters that explore the "language" of metadata as a means to promote discovery and enable access to information resources curated by many parties. We invite you to take full advantage of the DC-2015 conference to learn, share, and network with fellow experts and researchers. As with all DCMI events, participants are encouraged to be cordial and respectful towards one another, and to take opportunities to introduce themselves to new people, and include other conference-goers in conversations and social events. A friendly climate adds to everyone's conference experience.

Also, feel free to participate in the social media dimensions of the conference by:

- Following @dcmi15
- Tweeting using the hashtag, #dcmi15
- Adding yourself to the conference Lanyrd page <http://lanyrd.com/cqtxd>

Post-conference, participation in DCMI is an excellent way to continue to be involved in the community. The DCMI Annual Meeting (following the DC-2015 Conference on Friday) will be an opportunity to do strategic thinking about DCMI's directions – please join us! And please follow DCMI on Twitter (@dublincore) and consider becoming a DCMI Individual Member (see <http://dublincore.org/support/>).

Again, our thanks for your participation in the DC-2015 conference. And our special thanks to the Universidade Estadual Paulista, the Organizing Team, conference volunteers, and DC-2015 sponsors for making this event possible.
Enjoy the conference!

Eric Childress, Chair, DCMI Governing Board



DC-2015

Bem-vindo

Bem-vindo à DC-2015 em São Paulo, Brasil! Em nome da Iniciativa de Metadados Dublin Core (DCMI) é com prazer que me junto ao nosso anfitrião, a Universidade Estadual Paulista, e à Equipe de Organização, para vos dar as boas-vindas à Conferência e Encontro Anual Dublin Core DC-2015.

A Conferência internacional anual da Dublin Core e de aplicações de metadados estabeleceu-se como uma das principais conferências do mundo, reunindo investigadores, profissionais, implementadores e estudantes activos em metadados no sector cultural e sectores relacionados com aprendizagem. Este encontro marca a 15ª Conferência da série de conferências que começou em Tokyo em Outubro de 2001. Esta é a primeira vez que o evento acontece na América do Sul, e somente a segunda vez que decorre num país de língua oficial Portuguesa (DC-2013 aconteceu em Lisboa).

O tema deste ano, Metadados e acesso ubíquo à cultura, ciência e humanidades digitais, fala da necessidade de dados estruturados suportarem acesso ubíquo em toda a Web de forma a preservar os achados culturais de todas as culturas, na ciência e nas humanidades digitais. Esta necessidade é agora senso comum nos designers de sistemas de informação e implementadores. Com este fim, a DC-2015 contará com tutoriais, sessões especiais, artigos, relatórios de projectos, e posters que irão explorar a “linguagem” dos metadados como forma de promover a descoberta, e facilitar o acesso a recursos de informação disponibilizados por muitas organizações e instituições.

Convidamo-lo a aproveitar ao máximo a Conferência DC-2015, aprendendo, partilhando conhecimento, e realizando networking com especialistas e investigadores. Em relação aos eventos da DCMI os participantes são convidados a serem cordiais e a respeitarem os participantes; são ainda convidados a agarrar oportunidades de se apresentar a novas pessoas, e a incluir outros conferencistas em conversas e eventos sociais. Um clima amistoso acrescenta sempre algo à experiência de todos.

Convidamo-lo ainda a participar na dimensão dos media sociais da Conferência através dos canais:

- Seguindo @dcmi15
- Tweetando usando a hashtag #dcmi15
- Adicionado-se à página Lanyrd da Conferência <http://lanyrd.com/cqtxd>

Para terminar, referimos ainda a importância de participar nos eventos da DCMI. É uma forma excelente de continuar envolvido na comunidade. O encontro anual da DCMI (no seguimento da Conferência DC-2015, na sexta-feira) será uma oportunidade de realizar pensamento estratégico, reflectindo nas direcções de futuro da DCMI – junte-se a nós! Além disso, siga a DCMI no Twitter (@dublincore) e pense em tornar-se membro individual da DCMI (veja como em <http://dublincore.org/support/>).

Mais uma vez obrigada por participar na Conferência DC-2015. Deixamos os nossos agradecimentos especiais à Universidade de Estadual Paulista, à Equipe de Organização,

aos voluntários da Conferência, e aos patrocinadores DC-2015 por tornarem possível este evento.

Disfrute da conferência!

Eric Childress, Presidente da Direcção da DCMI.



Chair's Notes on the Program

This year, the *International Conference on Dublin Core & Metadata Applications* takes place under the general theme *Metadata and Ubiquitous Access to Culture, Science and Digital Humanities*. The exponential growth in the use of mobile devices has led to an equally exponential growth in information demand. We know well, as final consumers, the convenient use of these Apps in our mobile devices. This growth presents an enormous opportunity and a great challenge for the metadata community. That challenge must include as part of our research and development agenda the development of data models and metadata practices capable of responding with efficacy to this demand for information. The Linked Open Data movement can gain an enormous momentum in this era of ubiquity.

We are very happy to confirm, as expected, many submissions from our Brazilian colleagues. In fact, reaching the regional community of research and practice is one of the reasons why the DCMI conference is itinerant allowing it to challenge the local communities in the Americas, Europe, and Asia to “step up” and be present and to show their work to an international audience through Technical and Professional programs designed to span the domains of research and the practice that research informs.

This year the conference will have two keynote speakers. Paul Walk from Edina (United Kingdom) will deliver an address Wednesday titled “Exploiting the value of Dublin Core through pragmatic development”. Ana Alice Baptista, from the University of Minho (Portugal), will speak on Thursday about “Application Profiles and why the ‘how’ is important”.

The Technical Program is devoted to research papers and project reports in six sessions throughout the two Conference days—each session having the presentation and discussion of a mix of three papers and projects. On the first day, there will be two parallel Technical sessions in the morning, “Application Profiles and Ontologies” and “Studies on Metadata Practices”. In the afternoon, there are two Technical sessions—the first on the topic of “Reflections and developments on Application Profiles” and the second on “Metadata Migration”. In the morning of the second day, we have the session “Digital Repositories” and in the afternoon the session “Metadata Praxis”.

The first Professional session occurs in the afternoon of the first day and is titled “Linked data in the context of the cultural heritage”. In this session, researchers present projects committed to the publication of cultural heritage data as Linked Data while illustrating some of the problems that arose and the solutions found in making data appropriate for the Semantic Web. In the afternoon of the first day, the second session called “Schema.org Structured Data on the Web—An Extending Influence” will be comprised of a small number of scene setting presentations from those active in the evolution and application of schema.org, followed by discussion. In the afternoon of the second day, the Professional session titled “Recent Developments in Metadata for Research Data” will gather researchers that will report recent metadata activities that are relevant for the management and curatorship of the research data.

Full-day workshops in the pre- and post-conference training program will target professional interested in the approached themes. This year, two workshops will take place in parallel on the pre-conference day. One of them follows in the footsteps of its successful presentation last year at DC-2014 in Austin titled “Training the Trainers for Linked Data” in which participants will learn to clean, reconcile, enrich and publish metadata in the context of the Linked Open Data using freely available tools and data. In the other workshop titled “Development of Application Profiles”, the participants will use a specific method for development of a application profiles (Me4MAP). On the post-conference day, a workshop

titled “Elaboration of SKOS Controlled Vocabularies” will teach participants to develop short vocabularies in SKOS using a free tool for ontology edition called “Protégé”. On this same day, in parallel with this workshop, the Annual Meeting of the DCMI will take place, where important issues regarding the strategic planning for the organization will take place.

DC-2015 marks the 15th year of the DCMI International Conference. We celebrate these 15 years by opening up to a new continent in the south. This Conference takes place for the first time in South America—more specifically in São Paulo, Brazil. We hope this is a good opportunity for the information systems designers and implementers in South America to create a space for reflection and knowledge sharing, providing discussions that create new ties and new synergies for the future. May the conference held in the South bring new members to the DCMI community, making it richer and even more diverse.

Welcome to DC 2015!

Mariana Curado Malta,

CEISE/ISCAP - Polytechnic of Oporto, Portugal
& Algoritmi Research Center - University of Minho, Portugal

Silvana A. Borsetti Gregorio Vidotti,

Universidade Estadual Paulista (UNESP), Brazil



Notas dos responsáveis pelo programa técnico sobre o programa

Neste ano de 2015 a Conferência Internacional sobre o Dublin Core e Aplicações de metadados decorre sob o tema geral de Metadados e Acesso Ubíquo à Cultura, Ciência e Humanidades Digitais. O aumento exponencial da utilização dos dispositivos móveis tem como consequência também um aumento da procura de informação. Sabemos bem, como utilizadores finais que somos, da conveniência na utilização das Apps nos nossos dispositivos móveis. Este facto é uma enorme oportunidade e um grande desafio para a comunidade de metadados. Esta comunidade deverá ter na ordem do dia o desenvolvimento de modelos de dados e de práticas de metadados capazes de responder de uma forma eficaz a esta procura. O movimento de dados abertos ligados na Web poderá ganhar um enorme ímpeto nesta era da ubiquidade!

É com alegria que constatamos que tivemos muitas submissões por parte dos nossos pares Brasileiros. De facto, esta é também uma das razões pela qual a conferência é itinerante, ela desafia as comunidades locais nas Américas, na Europa, e na Ásia a tornarem-se presentes, e a mostrarem os seus trabalhos a uma audiência internacional através de programas técnicos e profissionais desenhados para expandir os domínios de investigação e a prática que a investigação informa.

A Conferência terá este ano dois oradores principais. Paul Walk da Edina (Reino Unido) que irá falar sobre “Explorando o valor de Dublin Core através do desenvolvimento pragmático”, e Ana Alice Baptista, da Universidade do Minho (Portugal) que irá falar sobre “Perfis de Aplicação e por que razão o “como” é importante”.

O programa técnico dedica-se a artigos de investigação e relatórios de projectos em seis sessões durante os dois dias da conferência – cada sessão tem a apresentação e discussão de um mix de três artigos e projectos. No primeiro dia teremos de manhã duas sessões técnicas em paralelo, “Perfis de Aplicação e Ontologias” e “Estudos sobre práticas de metadados”. De tarde teremos duas sessões técnicas seguidas, a primeira “Reflexões e desenvolvimentos sobre Perfis de Aplicação” e a segunda “Migração de metadados”. Na manhã do segundo dia teremos a sessão técnica “Repositórios Digitais” e à tarde a sessão técnica “Metadata Praxis”.

A primeira sessão profissional ocorre na tarde do primeiro dia, ela intitula-se “Dados ligados no contexto da herança cultural”. Nesta sessão, alguns investigadores irão apresentar projectos empenhados na publicação dos dados da herança cultural como Dados Ligados e mostrarão alguns dos problemas que surgiram e algumas das soluções que foram encontradas para tornar os dados apropriados à Web Semântica. Ainda na mesma tarde do primeiro dia da Conferência, a segunda sessão profissional intitulada “Schema.org – dados estruturados na Web, uma influência que se estende” irá abarcar um pequeno número de apresentações feitas pelas pessoas que têm estado activas no desenvolvimento e aplicação do Schema.org; estas apresentações serão seguidas de uma discussão. Na tarde do segundo dia da Conferência, a sessão profissional intitulada “Recentes desenvolvimentos em metadados para dados de investigação” trará investigadores que irão relatar actividades de metadados recentes que sejam relevantes para a gestão e curadoria dos dados de pesquisa.

Nos dias precedente e posterior à Conferência haverá um programa de formação com workshops que duram um dia completo, este programa tem como público alvo profissionais interessados nos temas abordados. Este ano irão decorrer duas workshops em paralelo no primeiro dia, uma delas no seguimento do sucesso que teve o ano passado na DC-2014 em Austin (Estados Unidos da América), ela será esta ano repetida com as actualizações necessárias. Com o título “Formar os Formadores de Dados Ligados”, os participantes, utilizando ferramentas e dados que estão disponíveis em utilização livre, irão aprender a limpar, reconciliar, enriquecer e publicar metadados no contexto dos dados ligados na Web. Na outra workshop, com o título “Desenvolvimento de perfis de aplicação” os participantes serão convidados a desenvolver um perfil de aplicação, aplicando um método para o desenvolvimento de perfis de aplicação (Me4MAP). No dia após a Conferência irá decorrer a última workshop “Elaboração de Vocabulários controlados em SKOS” onde os participantes irão aprender a desenvolver pequenos vocabulários na linguagem SKOS utilizando uma ferramenta livre de edição de ontologias, o “Protégé”. Neste mesmo dia ocorrerá, em paralelo com esta workshop, a reunião anual da DCMI onde se discutirão temas importantes de planeamento estratégico da organização.

Não queremos deixar de marcar os 15 anos da Conferência Internacional sobre o Dublin Core e Aplicações de metadados - celebramos estes 15 anos abrindo um novo continente ao Sul. Esta Conferência acontece pela primeira vez na América Latina, mais concretamente na cidade de S.Paulo, no Brasil. Esperamos que ela seja uma boa oportunidade para que desenhadores e implementadores de sistemas de informação da América do Sul possam criar um espaço de reflexão e de partilha de conhecimento, proporcionando discussões que criem novos laços e novas sinergias para o futuro. E que a conferência realizada a Sul traga novos membros para a comunidade DCMI, tornando-a mais rica e ainda mais diversa.

Seja bem-vind@ à DC 2015!

Mariana Curado Malta,

CEISE/ISCAP - Polytechnic of Oporto, Portugal

& Algoritmi Research Center - University of Minho, Portugal

Silvana A. Borsetti Gregorio Vidotti,

Universidade Estadual Paulista (UNESP), Brazil



ORGANIZING COMMITTEE

GENERAL CONFERENCE CHAIR

Placida L. V. Amorim da Costa Santos, Universidade Estadual Paulista (UNESP), Brazil

PROGRAM CHAIR

Flávia Maria Bastos, Universidade Estadual Paulista (UNESP), Brazil

TECHNICAL PROGRAM CHAIRS

Mariana Curado Malta, CEISE/ISCAP - Polytechnic of Oporto, Portugal & Algoritmi
Research Center - University of Minho, Portugal

Silvana A. Borsetti Gregorio Vidotti, Universidade Estadual Paulista (UNESP), Brazil

PROFESSIONAL PROGRAM CHAIR

Stuart A. Sutton, Dublin Core Metadata Initiative (DCMI), United States

UNESP HOSTING COMMITTEE

Flávia Maria Bastos, Universidade Estadual Paulista (UNESP), Brazil

Placida L. V. Amorim da Costa Santos, Universidade Estadual Paulista (UNESP), Brazil

Silvana A. Borsetti Gregorio Vidotti, Universidade Estadual Paulista (UNESP), Brazil

LOCAL ORGANIZING COMMITTEE

Laura Andrade, Universidade Estadual Paulista (UNESP), Brazil

Felipe Augusto Arakaki, Universidade Estadual Paulista (UNESP), Brazil

Silvana Fagundes, Universidade Estadual Paulista (UNESP), Brazil

Cristina Rosa Teixeira, Universidade Estadual Paulista (UNESP), Brazil

DCMI ANNUAL MEETING COMMITTEE

Joseph T. Tennis, University of Washington, United States

Paul Walk, EDINA, United Kingdom

Eric Childress, OCLC Research, United States

Michael D. Crandall, University of Washington, United States

Kai Eckert, Hochschule der Medien (Stuttgart Media University), Germany

Valentine Charles, Europeana Foundation, Netherlands

Thomas Baker, Sungkyunkwan University, Korea, & Dublin Core Metadata Initiative (DCMI),
Germany

Stuart A. Sutton, Dublin Core Metadata Initiative (DCMI), United States

TECHNICAL PROGRAM COMMITTEE

Leif Andresen, The Royal Library. National Library of Denmark, Denmark

Thomas Baker, Sungkyunkwan University, Korea, & Dublin Core Metadata Initiative (DCMI),
Germany

Ana Alice Baptista, Universidade do Minho, Portugal

Uldis Bojars, National Library of Latvia, Latvia

Dan Brickley, Vrije Universiteit Amsterdam

Joseph A. Busch, Taxonomy Strategies, United States

Eric Childress, OCLC Research, United States

Michael D. Crandall, University of Washington, United States

Jacques Ducloy, University of Lorraine, France

Gordon Dunsire, Independent Consultant, United Kingdom

Kevin Ford, Library of Congress, United States



Muriel Foulonneau, Public Research Centre Henri Tudor, Luxembourg
Anne Gilliland, Department of Information Studies, UCLA, United States
Carol Jean Godby, OCLC, United States
Jane Greenberg, Drexel University, United States
Corey A. Harper, New York University
Bernhard Haslhofer, University of Vienna, Austria
Eero Hyvönen, Aalto University, Finland
Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Netherlands
Masahide Kanzaki, Keio University Xenon Limited Partners, Japan
Tomi Kauppinen, Aalto University School of Science, Finland
Dean Blackmar Krafft, Cornell University Library, United States
Michael Lauruhn, Elsevier, United States
Akira Maeda, Ritsumeikan University, Japan
Mariana Curado Malta, CEISE/ISCAP - Polytechnic of Oporto, Portugal & Algoritmi Research Center - University of Minho, Portugal
Philipp Mayr, GESIS - Leibniz Institute for the Social Sciences, Germany
Eva M. Méndez, University Carlos III of Madrid, Spain
Steven J. Miller, University of Wisconsin-Milwaukee School of Information Studies, United States
William E. Moen, University of North Texas, United States
Peter E. Murray, LYRASIS, United States
Jin-Cheon Na, Nanyang Technological University, Singapore
Liddy Nevile, retired from paid work but active in metadata work, Australia
Adrian Ogletree, Drexel University, United States
Johan Oomen, Netherlands Institute for Sound and Vision, Netherlands
Jung-ran Park, College of Computing and Informatics, Drexel University, United States
Oknam Park, Sangmyung University, Republic of Korea, Korea, Republic Of
Susanna Peruginelli, Free lance library consultant, Italy
Magnus Pfeffer, Stuttgart Media University, Germany
Sarah Potvin, Texas A&M University Libraries, United States
Jian Qin, Syracuse University, United States
KS Raghavan, Centre for Knowledge Analytics & Ontological Engineering (KAnOE), PES Institute of Technology, India
Stefanie Ruehle, SUB Goettingen, Germany
Placida L. V. Amorim da Costa Santos, Universidade Estadual Paulista (UNESP), Brazil
Johann Wanja Schaible, GESIS - Leibniz-Institute for the Social Sciences, Germany
Jodi Schneider, INRIA Sophia Antipolis, France, France
Ryan Shaw, University of North Carolina at Chapel Hill, United States
Aida Slavic, UDC Consortium, United Kingdom
Shigeo Sugimoto, University of Tsukuba, Japan
Stuart A. Sutton, Dublin Core Metadata Initiative (DCMI), United States
Lars G. Svensson, Deutsche Nationalbibliothek, Germany
Hannah Tarver, University of North Texas Libraries, United States
Vassilis Tzouvaras, National Technical University of Athens, Greece
Annelies van Nispen, Eye Film Institute, Netherlands
Sherry L. Vellucci, University of New Hampshire, United States
Paul Walk, EDINA, United Kingdom
Laura Waugh, University of North Texas, United States
Andrew C. Wilson, Queensland State Archives, Australia
Oksana Zavalina, University of North Texas, United States
Marcia Lei Zeng, Kent State University, United States



TABLE OF CONTENTS

i-iii	Welcome Message from DCMI Chair (EN & PT)
iv-vii	Chairs Notes (EN & PT)
viii-ix	Conference Committees
x-xii	Table of Contents
xiii-xv	Author Index

Session 1A—Application Profiles & Ontologies

1-9	MOD: Metadata for Ontology Description and Publication <i>Biswanath Dutta, Durgesh Nandini & Gautam Kishore Shahi</i>
10-19	A DCAP to Promote Easy-to-Use Data for Multiresolution and Multitemporal Satellite Imagery Analysis <i>Isabelle Mougenot, Jean-Christophe Desconnets & Hatim Chahdi</i>
20-29	A DCAP for the Social and Solidarity Economy <i>Mariana Curado Malta, Ana Alice Baptista & Cristina Parente</i>

Session 1B—Studies in Metadata Practices

30-40	An Exploratory Analysis of Subject Metadata in the Digital Public Library of America <i>Hannah Tarver, Mark Phillips, Oksana Zavalina & Priya Kizhakkethil</i>
41-49	Exposing Library Holdings Metadata in RDF Using Schema.org Semantics <i>Myung-Ja K. Han, Timothy W Cole, Patricia Lampron, M. Janina Sarol</i>
50-62	Exploratory Analysis of Metadata Edit Events in the UNT Libraries' Digital Collections <i>Hannah Tarver & Mark Phillips</i>

Session 2—Application Profiles: Reflections & Developments

63-75	Evolution of an Application Profile: Advancing Metadata Best Practices through the Dryad Data Repository <i>Edward M. Krause, Erin Clary, Adrian Ogletree & Jane Greenberg</i>
76-86	Do We Need Application Profiles? Reflections and Suggestions from Work in DCMI and ISO/IEC <i>Eva M. Méndez & Liddy Neville</i>
87-94	Language-Acquisition Inspired Sustainability Modeling for Application Profiles <i>Emma Tonkin</i>

Session 3—Metadata Migration

95-111	Guidance, Please! Towards a Framework for RDF-based Constraint Languages <i>Thomas Bosch & Kai Eckert</i>
112-118	Metadata Quality Control for Content Migration: The Metadata Migration Project at the University of Houston Libraries <i>Andrew Weidner & Annie Wu</i>
119-128	Understanding Metadata Needs when Migrating DAMS <i>Ayla Stein & Santi Thompson</i>



Session 4—Current Developments in Metadata for Research Data

- 129-135 Dublin Core Usage for Describing Documents in Brazilian Government Digital Libraries
Diego José Macêdo, Milton Shintaku & Ronnie Fagundes de Brito
- 136-145 BEAM Repository: A Proposal for Family and Personal Repository
Rachel Cristina Vesu Alves, Ana Carolina Simionato, Felipe Augusto Arakaki, Paula Regina Ventura Amorim Gonçalves, Ana Paula Grisoto & Plácida Leopoldina Ventura Amorim da Costa Santos
- 146-157 The Use of Application Profiles and Metadata Schema by Digital Repositories
Morgana Carneiro Andrade & Ana Alice Rodrigues Pereira Baptista

Session 5—Metadata Praxis

- 158-169 A DDC Visual Interface for Metadata Exploration
Xia Lin, Michael Khoo, Jae-wook Ahn, Ceri Binding, Douglas Tudhope, Hilary Jones & Diana Massam
- 170-180 Leveraging SKOS to Trace the Overhaul of the STW Thesaurus for Economics
Joachim Neubert
- 181-189 The Linkable Neil Armstrong: Using BIBFRAME to Increase Visibility of Digital Collections
Carolyn Hansen & Sean Crowe

Posters (Peer Reviewed)

- 190-191 EZID: Easy Identifier and Metadata Management
John Kunze, Greg Janée & Joan Starr
- 192-194 Using Metadata for Interoperability of Species Distribution Models
Cleverton Ferreira Borba & Pedro Luiz Pizzigatti Correa
- 195-197 Interlinking Two Institutional KOS about Agroecology: Using LOD Agrovoc to Circumvent the Language Barrier in Identifying Terminological Intersections
Sophie Aubin, Pascal Aventurier, Ivo Júnior Pierozzi & Leandro Henrique Mendonça Oliveira
- 198-200 Dublin Core and CIDOC CRM Harmonization
Laís Carrasco & Silvana A. Borsetti Gregorio Vidotti
- 201-202 LD4PE: A Competency-Based Framework for DCMI's Professional Education and Training Agenda
Thomas Baker, Michael D. Crandall, Stuart A. Sutton & Marcia Lei Zeng
- 203-205 How Should We Teach Metadata? What Comparisons Between Job Ad and Classroom Trends Can Tell Us About Preparing LIS Students
Deborah Maron & Jacob Hill
- 206-208 The Sweetpotato Ontology
Vilma Rocio Hualla Mamani, Reinhard Simon, Robert Mwanga, Henry Saul Juarez Soto & Genoveva Rossel Montesinos
- 209-211 Advancing Materials Science Semantic Metadata via HIVE
Yue Zhang, Jane Greenberg, Adrian Ogletree & Garritt Tucker
- 212-213 Study of Adhesion between Dublin Core and Marc: Reviewing the Interoperability between UNESP and the National Library
Elizabete Cristina de Souza de Aguiar Monteiro, Elaine Parra Affonso & Ricardo César Gonçalves Sant'ana
- 214-217 Bringing a Small Archival Collection to Life on the Web: Remembering the Real Winnie
Sally Wilson & Marina Morgan



- 218-219 Metadata for Models Generated by openModeller
Agnei Silva, Cleverton Ferreira Borba & Pedro Luiz Pizzigatti Correa
- 220-222 Evolution of Dublin Core Metadata Standard: An Analysis of the Literature from 1995-2013
Felipe Augusto Arakaki, Plácida Leopoldina Ventura Amorim da Costa Santos, Rachel Cristina Vesu Alves
- 223-225 Adopting the Dublin Core Standard for Describing Open Scientific Data: The e-Quilt Prototype Experiment
Adriana Carla S. de Oliveira, Guilherme Ataíde Dias & Renata Lemos dos Anjos
- 226-228 Proposal of Application Profile for Digital Images for Libraries, Archives and Museums (DILAM) Conceptual Model
Ana Carolina Simionato & Plácida Leopoldina Ventura Amorim da Costa Santos

Best Practice Posters

- 229-230 Joá Archival Description Application Profile: Uma Proposta de Perfil de Aplicação Dublin Core e Encoded Archival Description a Partir da Normal Geral Internacional de Descrição Arquivística [ISAD(G)]
Diana Vilas Boas Souto Aleixo, Maria Elisabete Catarino, Ana Alice Rodrigues Pereira Baptista
- 231-233 Use of Dublin Core to Increase Public Transparency of Brazilian Senate's Bills Datasets
Fernando de Assis Rodrigues & Ricardo César Gonçalves Sant'Ana
- 234-235 Metadata Reuse to Populate an Institutional Repository: Procedures Applied in UNESP Institutional Repository
Silvana Aparecida Borsetti Gregorio Vidotti, Flávia Maria Bastos, Juliano Benedito Ferreira, Ana Paula Grisoto, Fabrício Silva Assumpção, Renata Eleutério da Silva, Vítor Silvério Rodrigues & Oberdan Luiz May
- 236-237 Metadata Extraction and Register for Enterprise Information Architecture in the Brazilian House of Representatives
Mariana Baptista Brandt
- 238-240 Gateway to Oklahoma History Case Study: Structured Data and Metadata Evaluation for Improved Image Resource Findability on the Web
Emily Ann Kolvitz
- 241-243 Data Harmonisation between National Library Board, National Archives and National Heritage Board of Singapore
Shan Shan Chan & Haliza Jailani
- 244-246 Arquitetura Semântica de Recuperação da Informação
Caio Saraiva Coneglian, Elvis Fusco & José Eduardo Santarem Segundo
- 247-248 Dublin Core: A Metadata Standard in the "3 Marys"
Ana Carla Cunha Nascimento, Rayssa Thaynara Madeira Correia, Márcio Bezerra Da Silva
- 249-251 Particle Physics Metadata Standards in the Tritium File Format
Kevin Wierman, Adrian Ogletree & Jane Greenberg
- 252-254 Use and Connect: Linked Open Data of the National Diet Library, Japan
Yoshikazu Nagai, Akiko Hashizume & Julie Fukuyama



AUTHOR INDEX

Affonso , Elaine Parra	212
Ahn , Jae-wook	158
Aleixo , Diana Vilas Boas Souto	229
Alves , Rachel Cristina Vesu	136, 220
Andrade , Morgana Carneiro	146
Arakaki , Felipe Augusto	136, 220
Assumpção , Fabrício Silva	234
Aubin , Sophie	195
Aventurier , Pascal	195
Baker , Thomas H.	201
Baptista , Ana Alice Rodrigues Pereira	20, 146, 229
Bastos , Flávia Maria	234
Binding , Ceri	158
Borda , Cleverton Ferreira	192, 218
Bosche , Thomas	95
Brandt , Mariana Baptista	236
Brito , Ronnie Fagundes de	129
Carrasco , Lais	198
Catarino , Maria Elisabete	229
Chahdi , Hatim	10
Chan , Shan Shan	241
Clary , Erin	63
Cole , Timothy W.	41
Coneglian , Caio Saraiva	244
Correa , Pedro Luiz Pizzigatti	192, 218
Correia , Rayssa Thaynara Madeira	247
Crandall , Michael D.	201
Cristina Aprente	10
Crowe , Sean	181
Da Silva , Márcio Bezerra	247
de Assis Rodriuges , Fernando	231
de Oliveira , Adriana Carla Silva	222
de Souza , Virgínia Miranda	212
Desconnets , Jean-Christophe	10
Dias , Guilherme Ataíde	222
dos Anjos , Renata Lemos	222
Dutta , Biswanath	1
Eckert , Kai	95
Ferreira , Juliano Benedito	234
Fukuyama , Julie	252
Fusco , Elvis	244
Gonçalez , Paula Regina Ventura Amorim	136



Greenberg , Jane	63, 209, 249
Grisoto , Ana Paula	136, 222
Han , Myung-Ja K.	41
Hansen , Carolyn	181
Hashizume , Akiko	252
Hill , Jacob	203
Hualla Mamani , Vilma Rocio	206
Jailani , Haliza	241
Janée , Greg	190
Jones , Hilary	158
Khoo , Michael	158
Kizhakkethil , Priya	30
Kolvitz , Emily Ann	238
Krause , Edward M.	63
Kunze , John	190
Lampron , Patricia	41
Lin , Xia	158
Macêdo , Diego José	129
Malta , Mariana	20
Maron , Deborah	203
Massam , Diana	158
May , Oberdan Luiz	234
Méndez , Eva	76
Monteiro , Elizabete Cristina de Souza de Aguiar	212
Montesinos , Genoveva Rossel	206
Morgan , Marina	214
Mougenot , Isabelle	10
Mwanga , Robert	206
Nandini , Durgesh	1
Nascimento , Ana Carla Cunha	247
Neubert , Joachim	170
Nevile , Liddy	76
Ogletree , Adrian	63, 209, 249
Oliveira , Leandro Henrique Mendonça	195
Phillips , Mark	30, 50
Pierozzi , Ivo Júnior	195
Rodrigues , Vítor Silvério	234
Sant'ana , Ricardo César Gonçalves	231, 212
Santarem Segundo , José Eduardo	244
Santos , Plácida Leopoldina Ventura Amorim da Costa	136, 220, 226
Sarol , M. Janina	41
Shahi , Gautam Kishore	1
Shintaku , Milton	129
Silva , Agnei	218
Silva , Renata Eleutério da	234
Simionato , Ana Carolina	136, 226



Simon , Reinhard	206
Soto , Henry Saul Juarez	206
Starr , Joan	190
Stein , Ayla	119
Sutton , Stuart A.	201
Tarver , Hannah	30, 50
Thompson , Santi	119
Tonkin , Emma	87
Tucker , Garritt	209
Tudhope , Douglas	158
Vidotti , Silvana A. Borsetti Gregorio	198, 234
Weidner , Andrew	112
Wierman , Kevin	249
Wilson , Sally	214
Wu , Annie	112
Yoshikazu , Nagai	252
Zavalina , Oksana	30
Zeng , Marcia Lei	201
Zhang , Yue	209



Application Profiles & Ontologies—Session 1A

MOD: Metadata for Ontology Description and Publication

Biswanath Dutta
DRTC, Indian Statistical
Institute
Bangalore, India
bisu@drtc.isibang.ac.in

Durgesh Nandini
B.C. Roy Engineering
College
Durgapur, India
durgeshnandini16@yahoo.in

Gautam Kishore Shahi Dept.
of IT, Birsa
Institute of Technology,
India
gautamshahi16@gmail.com

Abstract

Ontology is an important artifact of Semantic Web applications. Today, there are an enormous number of ontologies available on the Web. Even so, finding and identifying the right ontology is not easy. This is because the majority of ontologies are either not described or described with a general-purpose metadata vocabulary like Dublin Core. On the other hand, ontology construction, irrespective of its types (e.g., general ontology, domain ontology, application ontology), is an expensive affair both in terms of human resources and other infrastructural resources. Hence, the ideal situation would be to reuse the existing ontologies to reduce the development effort and cost, and also to improve the quality of the original ontology. In the current work we present an ontology metadata vocabulary called Metadata for Ontology Description and publication (MOD). To design the vocabulary, we also propose a set of generic guiding principles and a well-established methodology which take into account real concerns of the ontology users and practitioners.

Keywords: metadata, ontology metadata vocabulary, ontology publication, resource description, ontology reuse, ontology library, methodology, metadata design principles, semantic application

1. Introduction

Ontology (a formal, explicit specification of a shared conceptualization (Studer et al., 1998)) construction is an expensive affair both in terms of human and other infrastructural resources. One of the fundamental principles of ontology development is to look for existing ontologies to reuse (Dutta, B. et al., 2015) before deciding to construct one from scratch. In this context, a new type of library that stores ontologies, called an Ontology Library (Ding and Fensel, 2001; d'Aquin & Noy, 2012), plays a crucial role. The goal of an ontology library is to support users to search and retrieve ontologies for the purpose of reusing them. However, in our opinion as ontology practitioners, theoretically the idea of ontology reuse sounds appealing, but in practice it is not easy to implement. There are various reasons why reuse may not be easy to practice. For example, reuse, whether partially or in full, can happen only when there is a match between the user goal of using an ontology and the development goal of an existing ontology. Obrst, et al. (2014) has discussed many such concerns in the form of “what limits ontology reuse?” One of the possible concerns is highly relevant to the current work, i.e., how to find Mr. Right Ontology? To quote them:

...more than a simple registry of ontologies is needed – there must also be ways of organizing and annotating the ontologies with the appropriate metadata so that users can find the ontologies that match their requirements.

They further state that in addition to notions such as provenance (as captured by Ontology Metadata Vocabulary (OMV) (Hartmann, Jens et al., 2005), which is so far the only existing metadata vocabulary for ontology description), the metadata must include a wider range of features. For instance, metadata from a *development perspective* consists of information such as competency questions, ontological commitments and design decisions; metadata from an

implementation perspective consists of information for reasoning support, languages, rules, conformance to external standards and so forth.

The current work proposes an ontology metadata vocabulary, called Metadata for Ontology Description and publication (MOD). In designing MOD, we have considered the above recommendations of Obrst, et al. (2014) as well as recommendations made by various other ontology practitioners and users in the literature including d'Aquin & Noy (2012). The main contributions of the current work are as follows: proposes an easy to use and well-defined ontology metadata vocabulary MOD, which considers the real concerns of the practitioners and ontology users; proposes a set of *generic* guiding principles and a methodology for designing a metadata vocabulary.

The rest of the paper is organized as follows: section 2 discusses the current state of the ontology libraries. It provides answers to some of the following questions, such as how many metadata elements are used by existing ontology libraries? Do they use any standard vocabularies to describe the ontologies?; section 3 discusses MOD design principles as well as the methodology; section 4 discusses a set of top-level facets, describing the various perspectives of an ontology, that are defined to design the current MOD vocabulary; section 5 provides details of the MOD vocabulary; section 6 discusses some of the related state-of-the-art works. Finally, section 7 concludes the paper.

2. Ontology Metadata in Practice: The Current State of Ontology Libraries

In this section, we present the results of our study of the usage of metadata by the existing ontology libraries on the Web to describe and publish ontologies. Before we discuss the results, we will first briefly define an ontology library and discuss its purpose.

In general, an ontology library is a collection and organization of ontologies. The purpose of an ontology library is to allow users to search, browse, refer and evaluate ontologies for different tasks. The ontology libraries are generally classified into three broad categories: *ontology repository*, *ontology registry* and *ontology directory* (Debashis, N., 2014). We have identified a total of 13 such libraries on the Web. These include Bio-portal, DERI, OBO Foundry, ROMULUS, colore, etc. as shown in Table 1.

TABLE 1: Ontology libraries along with their number of metadata elements

Ontology Library	Number of Elements	Example Elements	Metadata Followed
Bio-Portal (https://bioportal.bioontology.org/)	30	Acronym, People, Number Of Properties, Status, Description	Partially OMV plus own defined elements
Colore (https://code.google.com/p/colore/source/browse/trunk/ontologies/approximate_point)	7	Source Path, File Name, Size, Rev, Author	None
DAML (http://www.daml.org/ontologies/)	12	Link, Description, Submitter, Point of contact, Submitter	None
DERI (http://vocab.deri.ie/)	4	Author, Terms, Last Update, Namespace URI	None
Maven (http://mvnrepository.com/artifact/edu.stanford.protege)	4	Artifact, Last Version, Popularity, Description	None
MISO (http://www.sequenceontology.org/)	6	Definition, Synonyms, DB Xref, Parent, Children	None
MMI (http://mmisw.org/)	22	Full Title, Contact Role, Syntax Format, Authority abbreviation, Contributor, Keywords	None
OBO Foundry (http://www.obofoundry.org)	12	Namespace, Current Activity, Help, Home, Documentation, Contact	None
ONKI (http://onki.fi/en/browser/)	11	Type, URI, Share, superordinate concepts, Coordinate concepts	None

Ontohub (https://ontohub.org/ontologies)	24	Project Name, Description, Institution, URL, task	Partially OMV plus own defined elements
ROMULUS (http://www.thezfiles.co.za/ROMULUS/)	35	Ontology Name, License Description, Project Domain, Creation date, DL expressivity, Number of classes, Number of individuals	Partially OMV plus own defined elements
Schemapedia (http://datahub.io/dataset/schemapedia)	4	Subject, Property, Source	None
SHOE (http://www.cs.umd.edu/projects/plus/SHOE/onts/)	4	Id, Version, Description, Contact	None

To understand the state-of-the-art practices and the metadata usages among the ontology libraries, we have studied each of these 13 libraries thoroughly and have noted the metadata elements they use. We have also tried to find information on whether any of these libraries follow a metadata standard. A consolidated result of our study based on the above parameters is presented in Table 1.

It can be seen from the above Table 1 that except three libraries, namely, Bio-portal, Ontohub and ROMULUS, none of the other libraries use a metadata standard or controlled vocabulary system in describing the ontologies. These three libraries partially use a metadata vocabulary called Ontology Metadata Vocabulary (OMV). In addition to OMV metadata elements, these libraries also use additional self-defined metadata elements. Zubeida and Keet (2013) have observed that OMV is not sufficient for an extensive and descriptive list of metadata for the ontologies. This deficiency in ontology metadata vocabulary may create an obstacle in ontology reuse. It can be further seen from the above table that the usage of a number of metadata elements varies from library to library. The majority of the libraries (70%) are found to be using 15 or fewer than 15 elements. This indicates that the metadata set should not be too large.

By analyzing the above libraries and their metadata, we have also observed that different terms are used in describing similar information in different libraries. For instance, the majority of the libraries have used the term “author” to capture the author information of an ontology, while some of the libraries have used the term “creator” (e.g., ROMULUS). This occurs when we do not use any standard or controlled vocabulary system. The practice of using ad hoc solutions creates obstacles in achieving interoperability among the ontology libraries.

Given the above observations, we have designed MOD as a controlled metadata vocabulary system that can be used by the community. We have tried to provide a minimal set of elements, but keep the essential elements that would be needed to describe an ontology and support ontology reuse.

3. MOD Approach

The MOD approach involves two crucial components: guiding principles and methodology. These are discussed in the following.

3.1 Guiding Principle

To design MOD, we developed some generic principles that acted as guidelines for us in the process of creating the vocabulary. The principles are important to assure the consistency and effectiveness of the vocabulary. The principles are:

1. *Principle of brevity*: The vocabulary should consist of a minimal set of elements maintaining balance between *necessity and sufficiency*.
2. *Principle of clarity*: The metadata elements must be well defined and clear descriptions should be provided.
3. *Principle of simplicity*: The vocabulary should be easy to use.

4. *Principle of authority*: The vocabulary design should be based on a sound methodology in the sense that the inclusion of terms in the vocabulary are justified.
5. *Principle of standardization*: The element names should be standardized. To confirm the standardization, the individual elements should be mapped with the existing standard vocabularies.
6. *Principle of extensibility*: The vocabulary should be extensible.
7. *Principle of usability*: The vocabulary should support the reuse of the described resources. In other words, the vocabulary should allow the creators/developers to highlight the usage and the quality of the resources in a well-defined manner.
8. *Principle of interoperability*: The vocabulary should be interoperable. It should conform to the major knowledge representation languages currently in use for Semantic Web (Berners-Lee, et al., 2001) applications.

3.2 Methodology

To build up MOD, we have used a two-way approach: Top-down approach and Bottom-up approach as discussed below.

Top-down approach

The top-down approach involves looking at the “big picture” of the metadata vocabulary. This is accomplished by defining the top-level facets conceiving the various aspects of the resource to be described. In the current work, the primary resource is an ontology. After defining the aspect, each aspect has to be further analyzed and narrowed down to define the various classes. So the top-down approach proceeds from an abstract level to a concrete level. A further explanation of this step, including the various top-level facets, is contained in section 4.

Bottom-up Approach

The bottom-up approach involves studying and identifying the properties of a resource for search and discovery to facilitate their effective reuse. This is accomplished by analyzing users' ontology search behavior, search criteria and parameters. The extracted properties are further associated with the classes defined in the top-down approach. So the bottom-up approach proceeds from a concrete level to an abstract level.

To understand the users' search behavior, search criteria and parameters, we have conducted a survey. For this, we have used an open-ended questionnaire as a tool. We circulated the questionnaire through email to people who use or deal with ontologies on a regular basis. Participation consisted of researchers and practitioners with diverse educational backgrounds including library and information science, computer science, philosophy, linguistics, etc. The participants were from various countries like India, Italy, Bangladesh, Palestine, etc. After receiving the replies, we have analyzed them and have extracted the key requirements in terms of metadata elements as discussed below.

Two specific questions were asked to the participants. These are:

- (1) *How do you search an ontology on the Web or in an ontology library?*
- (2) *When you search for an ontology, what is the information you look for before deciding to refer/ consult/ download it?*

We originally sent the questionnaire to a total of 18 people, out of which 12 people responded. As it was an open ended questionnaire, the responses were descriptive. Each of the responses consisted of multiple sentences (aka statements). Each sentences reflect the various actions and concerns of the participants in context of ontology search and retrieval. Table 2 lists the most frequently replied statements. The keywords of the statement have been *italicized*. MOD accommodates all of the essential and most frequently used keywords. These keywords have been

compiled and framed to form the elements of MOD. These responses have not only provided sufficient input for deciding the metadata elements of MOD, but have also provided a potential foundation to outline the multi-faceted approach to the metadata in the early stages of its development.

TABLE 2: Ontology user responses

<p>Statement 1: I look at the ontology descriptors like <i>domain details, number of classes, properties, tools used</i>.</p> <p>Statement 2: I look for <i>representations languages</i> while downloading an ontology.</p> <p>Statement 3: I look for SPARQL query file, if any.</p> <p>Statement 4: I would like to see 'user reviews' with these ontologies, so that I can save a lot of time in understanding the quality of the ontology.</p> <p>Statement 5: I prefer to have a <i>documentation/ information about the methodology</i> followed to develop ontology, it will be an additional advantage.</p> <p>Statement 6: I remain curious about the following facts: <i>top classes, number of classes and class definitions</i>.</p> <p>Statement 7: I look for <i>types and number of relations</i>.</p> <p>Statement 8: I look for <i>number of entities and description</i> about each of them.</p> <p>Statement 9: I look for whether I can export the whole or part of the ontology, also look for the <i>languages and formats</i> to export.</p>	<p>Statement 10: I look for whether the <i>ontology visualization</i> feature is supported.</p> <p>Statement 11: I look for the <i>date on which the ontology was created</i>.</p> <p>Statement 12: I look for if the ontology was created manually or through some kind of corpus mining, i.e. some information regarding <i>how the ontology was created</i>.</p> <p>Statement 13: I look for the <i>person or organization</i> that has developed the ontology.</p> <p>Statement 14: I look at the <i>classes and properties</i> of the ontology as they are very important in scrutinizing a basic evaluation of ontology; especially in those cases where I am searching for ontology of a known field or domain.</p> <p>Statement 15: I usually search ontology by <i>topic</i> and then see the <i>relevant classes</i>. Sometimes title does not reflect the relevant ontology classes.</p> <p>Statement 16: Before selecting an ontology to download, I make sure it is in OWL, because of my familiarity with this <i>ontology language</i>.</p>
---	---

4. Top-level Facets

Following the top-down approach, as discussed above, we have derived a set of top-level facets for the ontology vocabulary. The top-level facets provide a high-level schema of the ontology metadata vocabulary expressing the various aspects of an ontology. These facets are further analyzed to define the classes of MOD discussed in section 5.

To derive the top-level facets, we treat the ontology as a study of subject. In other words, an ontology is at the center of our study. We have studied and analyzed an ontology from multiple perspectives. There are a total of seven aspects that have been identified as follows:

- *General*- an abstraction of the general aspects of an ontology, for instance, the ontologies, ontology type, etc.
- *Ontology Coverage*- an aspect that defines the domain (*a domain is any area of knowledge or field of study that we are interested in or that we are communicating about that deals with specific kinds of entities* (Giunchiglia and Dutta, 2011, Giunchiglia, et al., 2014)) and scope of an ontology.
- *Authority*- describes the agents, like organizations, that own and are responsible for the ontology.
- *Rights*- describes the rights and licenses of an ontology.
- *Environment*- defines the environment in which an ontology has been built, for instance, the tool that is used to build an ontology, the level of formality, and the syntax followed.
- *Action*- an aspect highlighting the applications where an ontology is being applied or used, such as in a project.
- *Preservation*- describes the low level-features of an ontology, for instance, ontology storage, file format, etc.

It can be seen from the above descriptions that each of these aspects is complex in nature. We have further analyzed these aspects and have derived the basic classes of MOD as discussed in the following section.

5. MOD Metadata Model

MOD, the metadata vocabulary, consists of 64 elements. These elements are expressed in terms of Classes, Object properties and Data properties. There are 15 classes including two subclasses, 18 object properties and 31 data properties. The further descriptions on the classes and properties are provided below. In the successive sections we also discuss the various controlled vocabularies that are used to create and standardize the MOD vocabulary.

To make the MOD vocabulary interoperable and conform to the major representation languages currently being used for the Semantic Web applications, we have expressed MOD using OWL. The ontology is available at <http://www.isibang.ac.in/~bisu/>.

5.1 Classes

MOD consists of 15 classes (*a class is a collection of things sharing common attributes*) presented in Table 3 along with some exemplary class instances. Classes are important in metadata vocabulary as they are required to represent and support the reuse of ontologies (Hartmann, et al., 2005). The classes are derived by analyzing the top-level facets described above. For instance, the top-level facet *Authority* refers to a person or an organization who created and/ or who exercises control over an ontology, an ontology document, etc. In MOD both *Person* and *Organization* are considered as classes and grouped under a general class *Agent*. Similarly, by analysing the facet *Environment*, we have derived the classes like *OntologyTool*, *OntologyLanguage*, and *OntologySyntax*. In a similar fashion, we have analysed all the top-level facets and have derived the following classes shown in Table 3. The classes are presented in the table with the corresponding facets.

TABLE 3: MOD ontology classes

Top-level facet	Class Name	Example of Class Instances
General	Ontology	Space ontology, Food ontology, Fishery ontology,
Authority	Agent Subclass : Organization Subclass : Person	Organization related with the ontology and the person associated with it.
Right	License	Creative Commons, GNU Free Documentation License, GNU General Public License
Scope/Coverage	Domain	Genes, Space, Medicine, Protein
	Ontology type	Application Ontology, General Ontology, Core Reference Ontology
Action	Project	Smart city, Mobility
	Methodology	METHONTOLOGY, YAMO
Environment	Ontology design tool	OntoEdit, Protégé, TopBraid composer
	Ontology design language	RDFS, OWL
	Ontology design syntax	Notation3, Turtle, RDF/XML
Preservation	File Format	.rdf, .gaf
	Level Of Formality	Dictionary, Glossary
	Knowledge Representation Formalism	Frame, Description Logics, First Order Logic.

5.2 Object Property

Object property is a property that connects two resources belonging to two different, or the same, classes. MOD consists of 18 object properties including *creator*, *contributor*, *endorsedBy*, *evaluatedBy*, *module*, *formalityLevel*, *subject*, *usedIn*, etc. Each object property is defined with its domain and range. For instance, the object property *creator* has a domain class *Ontology* and a range class *Agent*. An object property can have more than one domain and range.

5.3 Data Property

Data property is a property that connects a resource to a data type. The data types are literals. MOD consists of 31 data properties, out of which 21 properties are directly the properties of an ontology resource. The other ten properties are the properties of other related resources, for instance, an Agent. Some of the data properties are: *name*, *acronym*, *identifier*, *noOfClasses*, *noOfProperties*, *noOfAxioms*, *naturalLanguage*, *lastUpdated*, *version*, etc. Each data property is specified with its domain and range. For instance, the data property *noOfClasses* has a domain class *Ontology* and a range *integer*. A data property can have more than one domain.

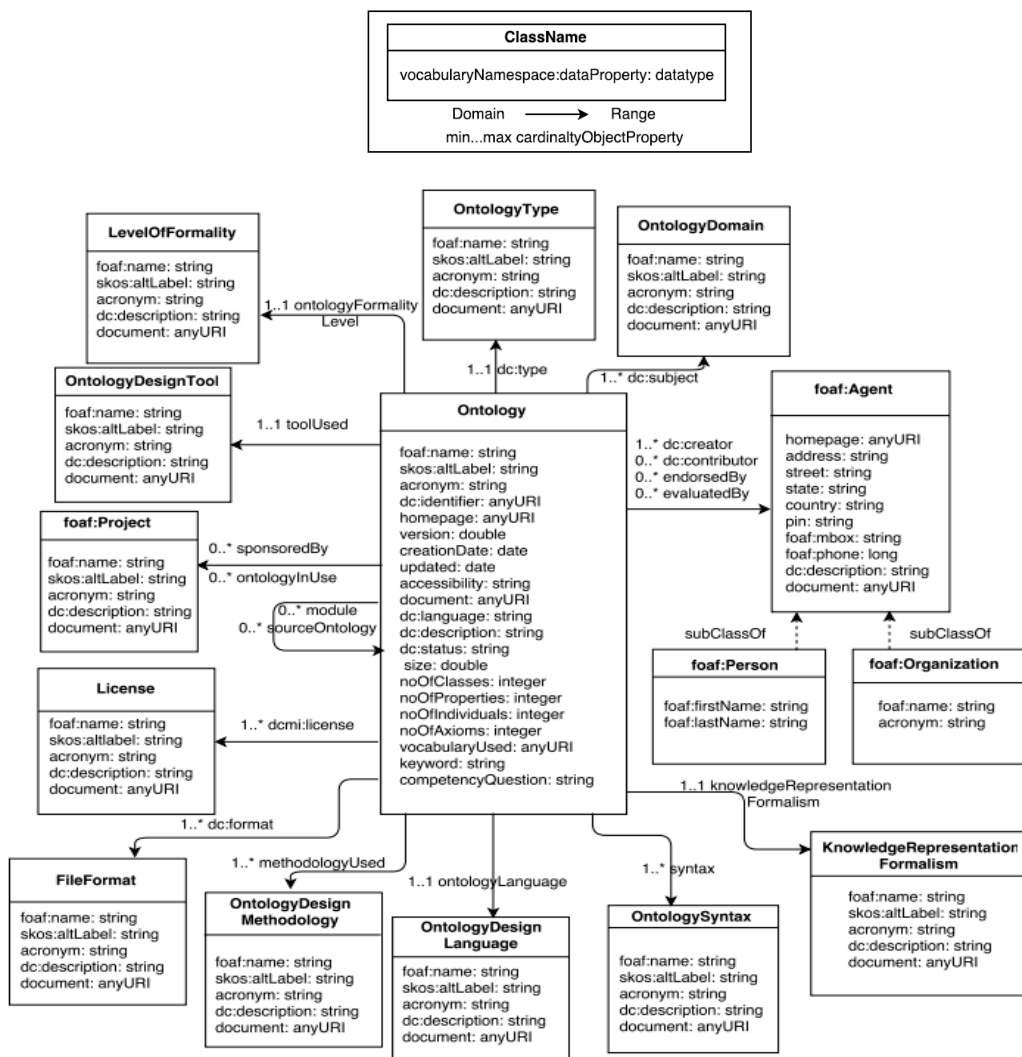


FIG. 1: MOD Overview

5.4 Vocabularies Used

The MOD terms are further standardised by using equivalent terms that are available in the existing metadata standards. Some of the metadata standards that we have used for this purpose are Friend Of A Friend (FOAF) (FOAF, 2014), Dublin Core (DC) (DCMI, 2015), and Simple Knowledge Organization System (SKOS) (SKOS, 2009). This way, MOD not only standardizes the vocabulary, but it also becomes part of the global initiative. This approach also ensures interoperability among the software programs.

Figure 1 above provides an overview of MOD vocabulary in terms classes, data properties and object properties including the constraints on the primary class Ontology. In the figure the prefixes represent the vocabulary namespace URIs. For dc, the namespace URI is <http://purl.org/dc/elements/1.1/>, for dcmi, the URI is <http://purl.org/dc/terms/>, for foaf, the URI is <http://xmlns.com/foaf/0.1/>, for skos, the URI is <http://www.w3.org/2004/02/skos/core#>.

6. Related Work

Here we will briefly discuss the related metadata standards, especially those that are relevant to the Semantic Web. DC Schema is a vocabulary consisting of a set of terms which can be used for describing web resources (video, images, web pages, etc.), as well as physical resources such as books, magazines, proceedings, journals, CDs, etc. Dublin Core has two sets of metadata, namely, unqualified DC (core elements) and qualified DC. FOAF provides a standard vocabulary to describe people, their activities and their relations to other people and objects. Anyone can use FOAF to describe him or herself. The Organization Ontology (Org, 2014) is a core ontology for organizational structures. It aims to support linked data publishing of organizational information across a number of domains. Its design goals are to allow domain-specific extensions to add classification of organizations and roles, as well as extensions to support neighboring information such as organizational activities. VOID (2011) is an RDF (Resource Description Framework) (RDF, 2014) vocabulary and a set of instructions. It enables the discovery and usage of linked-data sets. RDF Data Cube (2014) Vocabulary provides a means to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts using an RDF standard. It is a core foundation which supports extension of vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional data sets.

The above-discussed standards are related to our work in that they are metadata standards to describe the Web resources and are relevant for the Semantic Web applications. However, there is only one work that is very closely related to our work called Ontology Metadata Vocabulary (OMV). It provides a vocabulary for describing the ontologies. The basic differences between MOD and OMV are: MOD provides a minimized and well-defined set of metadata elements, which confirms the *principle of brevity* and *principle of clarity*. MOD elements are mapped and standardised with the other Semantic Web metadata standards. In other words, MOD reuses the existing metadata ontologies, which confirms the *principle of interoperability*. Overall, MOD is a well-guided, refined, easy-to-use standard ontology metadata vocabulary.

7. Conclusion

Metadata is instrumental in finding any kind of resources, whether they are print materials or electronic objects like ontologies, webpages, books, images, audio, video and so forth. Not only does metadata play a role in finding the resources, but can support in decision making to reuse the resources. In this context the current work has significance. MOD can be implemented by ontology libraries, and in general by Web developers, to make an ontology searchable and reusable. In our future work, we plan to pursue the use of MOD in the context of ontology libraries

Acknowledgement

We wish to thank Prof. Fausto Giunchiglia and Prof. Cristina Pattuelli for reviewing the work and providing valuable suggestions. We wish to thank the anonymous reviewers for their time in evaluating our paper patiently and providing critical comments and suggestions, which we have incorporated into this paper. We also thank Prof. A. R. D. Prasad and Prof. Devika P. Madalli for their constant support in conducting the current work.

References

- Studer, R. and Benjamins, V.R., Fensel, D. (1998). Knowledge engineering: Principles and methods. In *Data and Knowledge Engineering*, 25 (1-2), 161-197.
- Ding, Y., Fensel, D. (2001). Ontology library systems. The key to successful ontology reuse. In: *First Semantic Web Working Symposium*, 93–112.
- Dutta, B., Chatterjee, U. and Madalli, D. P. (2015). YAMO: Yet Another Methodology for Large-scale Faceted Ontology Construction. *Journal of Knowledge Management*. 19 (1), 6 – 24.
- d'Aquin, M., Noy, N.F. (2012). Where to publish and find ontologies? A survey of ontology libraries. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11, 96–111.
- Obrst, et. al. (2014). Semantic web and big data meets applied ontology. *Applied Ontology*, 9, 155-170.
- Hartmann, Jens et al. (2005). Ontology metadata vocabulary and applications. In *Proceedings of OTM Workshop (2005)*, LNCS 3762, 906–915.
- Debashis, N.. (2014). Ontology and ontology libraries: a critical study. In *Master's Dissertation (carried under the supervision of Biswanath Dutta)*. Bangalore, India: DRTC, Indian Statistical Institute, 10-49.
- Zubeida C. Khan and Keet, C. M. (2013). The foundational ontology library ROMULUS. *Model and Data Engineering*. LNCS, 8216, 200-211.
- Berners-Lee, Tim, Hendler, James and Lassila, Ora. (2001). The Semantic Web. *Scientific American*, 29-37.
- Giunchiglia, F. and Dutta, B. (2011). DERA: a Faceted Knowledge Organization Framework. Available at: <http://eprints.biblio.unitn.it/archive/00002104/>
- Giunchiglia, F., Dutta, B. and Maltese, V. (2014). From Knowledge Organization to Knowledge representation. *Knowledge Organization*, 41, 1, 44-56.
- OWL. (2004). OWL Web Ontology Language Overview. Retrieved July 9, 2015, from, <http://www.w3.org/TR/owl-features/>.
- FOAF. (2014). Vocabulary Specification 0.99. Retrieved July 9, 2015, from <http://xmlns.com/foaf/spec/>.
- DCMI. (2014). Dublin Core Metadata Element Set, Version 1.1. Retrieved July 9, 2015, from <http://dublincore.org/documents/dces/>.
- SKOS. (2009). SKOS Simple Knowledge Organization System Reference. Retrieved July 9, 2015, from <http://www.w3.org/TR/skos-reference/>.
- Org. (2014). The organization ontology. Retrieved July 9, 2015, from <http://www.w3.org/TR/vocab-org/>.
- VoID. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved July 9, 2015, from <http://www.w3.org/TR/void/>.
- RDF. (2014). RDF 1.1 Concepts and Abstract Syntax. Retrieved July 9, 2015, from <http://www.w3.org/TR/rdf11-concepts/>.
- Data Cube. (2014). The RDF Data Cube Vocabulary. Retrieved July 9, 2015, from <http://www.w3.org/TR/vocab-data-cube/>.

A DCAP to Promote Easy-to-Use Data for Multiresolution and Multitemporal Satellite Imagery Analysis

Isabelle Mougenot
Espace Dev UM,
France
isabelle.mougenot
@univmontp2.fr

Jean-Christophe Desconnets
Espace Dev IRD,
France
jeanchristophe.desconnets
@ird.fr

Hatim Chahdi
Espace Dev IRD & LIPN
UMR CNRS,
France
hatim.chahdi@ird.fr

Abstract

Satellite imagery can be exploited for any number of thematic analyses for Earth observation purposes. Characterization activities using remotely acquired data are currently made complicated by different limitations relating to, as an example, the meaningful mapping between multi-sensor data or the adding of the geospatial context to satellite information. We argue that describing satellite images through a metadata application profile may leverage capabilities to promote easy-to-use data for further in-depth thematic analysis. Accordingly, an application profile conforming to the Dublin Core application profile (DCAP) guidelines and designed for Earth observations (EO) is being developed. More specifically, we discuss RDF-compliant machine-processable aspects of the EO application profile (EOAP) in terms of the DCMI Description Set Profile (DSP) model. Additionally, a methodological approach to represent a DSP model using UML profiling activities is proposed.

Keywords: metadata; Dublin Core application profile; Earth observation, satellite imagery; semantic web standards; UML metamodeling

1. Introduction and motivations

Earth observing satellite imagery provides various datasets at different spatial, spectral and temporal resolutions. Each of these datasets can provide a complementary view that can improve assessments on the observed objects. The technical diversity of satellite sensors as well as their increasing number allows images to be considered at an unprecedented volume of data, richer and precise enough to deliver novel insights, such as a novel understanding of ecosystems dynamic or the monitoring of environmental changes at a local scale. The main objective of our DCAP is to integrate data with different spatial, spectral or temporal characteristics in an appropriate way, to gain more information that can be obtained from each individual sensor.

The increasing number of remotely sensed images as well as their large-scale distribution are the first impediments for data integration. Additionally, images are the results of numerous parameters, from technical characteristics of imaging sensors to atmospheric effects that limit capacities for systematic observations at various levels. Moreover, image-based data and their associated metadata are recorded in numerous file formats, such as GeoTIFF or JPEG 2000 that all have specific ways of describing content-based images.

In this context it is critical to simplify efficient image-based data access and query processing to provide accessibility to a variety of expert and non-expert users in remote sensing. Consequently the main aim is to document image-based data as well as additional data using metadata standards. Produced documents directly allow answering the query without consulting the data itself. For this purpose we have developed a metadata application profile according to the Dublin Core application profile (DCAP) guidelines (Nilsson, 2009). This application profile is named EOAP (Earth Observation Application Profile) (Desconnets, 2014) and is designed to benefit from metadata standards interoperability and linked open data principles for data sharing

on the web. Moreover, EOAP offers a descriptive framework that is flexible and extensible enough to adapt to numerous environmental uses cases as well as different viewpoints of users. The key objective is to ensure the use of the most comprehensive coverage of precise and accurate image-based data so that any environmental issues can be well addressed.

The current initiative is taken in the context of the GEOSUD research project¹. GEOSUD aims to promote increased access and use of satellite imagery to the French public. In particular the main objective of GEOSUD is to provide various facilities to effectively access shared image based data and processing tools that enable data retrieval, visualization and higher level analysis. GEOSUD was motivated by the lack of use of spatial data to help control natural environments and sustainable resources within land policy-making and institutional settings. Therefore a national spatial data infrastructure (Kazmierski, 2014) was developed to improve access to Earth observation data, particularly that of high and very high resolution satellite data. A suite of geospatial web services offers interoperable access to images provided by different satellite data suppliers as well as some image processing facilities.

The application profile we have developed will be a core component of this infrastructure in the near future and will be used to enable more advanced search analyses of Earth observation data. We therefore describe how EOAP may be used by referencing appropriate user scenarios.

Furthermore, we consider the application profile model for Earth observation as a domain specific language (DSL) [Fowler, 2010] and the constraint language DSP as a metamodel, which is more likely to permit the building of such a language. We draw on RDF and RDFS languages, UML profile and RDF metamodels using Meta-Object Facility (MOF) to build the DSP model for Earth observation.

The manuscript is structured in five sections. Following the introduction, Section 2 describes the diversity of datasets across imaging sensors and shortly introduces appropriate metadata standards that meet the variety of requirements to describe Earth observation data. Section 3 describes the modeling activities at different abstraction levels to work towards building a description set model for Earth observation resources. This model is RDF-compliant and conforms to the constraint language DSP. Section 4 provides illustrations about realistic use cases with specific needs related to multi-temporal and multi-resolution remotely sensed data that cover most significant functional aspects of the application profile under development. Finally, the last section draws preliminary conclusions and provides prospects about decisions made to implement the Earth observation application profile.

2. Background

2.1. Datasets

The satellite images that we want become accessible, via the GEOSUD spatial data infrastructure (SDI), are from different satellite sensors which acquire high and very high resolution images. Each year, these datasets must provide a high-resolution coverage of the entire national territory (5 meters/pixel) and in a second step a very high-resolution coverage (1.5 meters/pixel). In addition, in order to analyze the seasonal functioning of ecosystems and territories, time series with high frequency acquisition from medium resolution sensors, should also be available. Finally, the on-off scheduling of the acquisition of very high-resolution images via a direct receiving antenna is also planned. To meet these goals, several satellite sensors are used: Landsat 8 for medium resolution images, SPOT5 and Rapid Eye for high-resolution images, PLEIADES and SPOT6 for those with very high resolution. Established in 2011, the GEOSUD SDI is expected to acquire, each year, about 600 images, estimated at a volume of 1 to 2 TB / year for high resolution products and about 12 Tbytes / year for the very high resolution products.

¹ <http://www.equipex-geosud.fr>

² Clearcut: refers to a mode of forestry development through cutting down of all trees of a parcel

Depending on the image processing chain, the images are distributed in different formats such as GeoTiff or JPG2000 and, at different processing levels: projected raw images, radiometrically and geometrically corrected images, according to different encoding levels: 8 bits, 12 bits or 16 bits. Depending on the resolution and, the encoding format, an image can reach up to 50 MB for a medium resolution image (Landsat8), 2 GB for a high resolution image in JPG2000 format, even 15 GB for an equivalent image in GeoTiff format. Finally, these images are made accessible, searchable and downloadable via a set of web services and user applications.

2.2. Metadata framework

Different standards may be, general or dedicated to a particular discipline structure metadata. Regarding spatial data, we first want to mention the ISO 19115 (ISO, 2003) standard for geographic information, ISO 19115-2 (ISO, 2009) dedicated to gridded data and the specialization of O & M (Observation and Measurement) specification which proposes, among other things, elements to describe sensors characteristics and acquisition conditions (Gaspéri, 2012).

To better describe the satellite images, we chose among the metadata elements proposed by the various standards mentioned above. Moreover, we used the Dublin Core elements as common core elements such as title, creator, or coverage. The ISO 19115 and 19115-2 standards provide specific descriptors to the inherent spatial dimension to our datasets, such as the description of the spatial reference system associated with the location of the image (ISO19115: MD_ReferenceSystem), or the description of intrinsic characteristics associated with matrix structure of the image (e.g. ISO19115: MD_SpatialRepresentation). ISO 19115 and ISO 19115-2 also provide elements that characterize the sensor used to acquire the image (MI_Platform, MI_Instrument). Finally, the O & M for image description brings elements relating to acquisition conditions which are essential to pre-process the images after their acquisition.

A potential reproach to metadata standards is that they have been designed independently of each other and thus are not able to meet all information needs. This is especially true in our context in which the applications planned around images should cover a broad spectrum of functionalities: discovery, location, consultation, processing and, archiving. In fact, their implementation requires the contribution of individual standard. Based on this background, the definition of an application profile is relevant and able to offer a description framework both constrained but interoperable. The application profile built for satellite imagery (Desconnets, 2014) is based on the Singapore framework and application profile methodology named DCAP (Dublin Core Application Profile). Specifically, our work has focused on the definition of a

Description Set Profile, a structural model that completes the DCAM model to provide a prescriptive framework for the construction of an application profile (Nilsson, 2009). Thus, the application profile can be seen as a model that does not prescribe the data of interest, which are the satellite images, but the metadata elements which describe these datasets. The objective is both to reduce time and cost of datasets consultation and facilitate the management of big, heterogeneous and distributed data sources, such as the satellite images are. Their consultation is based on instances of DSP, i.e. metadata sets. In the next section, we will enlarge the implemented approach for the construction of the DSP model. The focus is given to modeling and metamodeling approaches.

2.3. RDF-compliant DSP to maximize reuse

We investigated the potential of using RDF language (Hayes, 2004) to build the DSP model. RDF is a W3C standard for encoding metadata, datasets and vocabularies on the web. Overall, we are giving top priority to release metadata instances in an open format on the web and we are considering this as a far better means of data exchange and sharing within the context of Linked Open Data (LOD) (Warren, 2014). DCMI provides some guidelines for encoding DSP specification in the RDF/XML concrete syntax (CWA15248, 2005). Additionally, many metadata standards including DCTERMS (DCES, 2012) are represented in a RDF serialization format, as

e.g. RDF/XML or N3. Similarly controlled vocabularies, as for example Geonames (Ahlers, 2013) or TGN (Getty Thesaurus of Geographic Names), are also available in RDF formats to guarantee open use.

However RDF and RDF Schema (RDFS) formalisms are built on a number of language primitives that are not always in line with the requirements drawn up for the constraint language DSP. In particular, RDF only provides a construct for declaring binary properties. Consequently, representing non-binary relations is a well-known issue, since n-ary relations arise quite commonly during modeling activities. A DSP model is intended to represent the overall structure of a metadata description set by means of constraints that apply either on resources described, properties used or values that may be given with respect to the properties. In this direction a DSP model is built using the notions of description template and statement template that define the valid skeleton of a description and a statement, respectively. A DSP is then a collection of description templates (*DescriptionTemplate*), which in turn are collections of statement templates (*StatementTemplate*). At the same time this notion of collection involves three entities, namely *DescriptionTemplate*, *Property* and *Constraint* and reveals a complex constraint that requires a ternary relation. A UML (Unified Modeling Language) class diagram (Rumbaugh, 1991) for specifying such collections is introduced (FIG. 1) and describes the class *DescriptionTemplate* associated with the classes *StatementTemplate*, *Property* and *Constraint* by means of an n-ary relationship. *StatementTemplate* is represented as an association class and appears as a class linked to the association with a dashed line. *DescriptionTemplate* contains a reference to *StatementTemplate*, which in turn contains a first reference to the class *Property* and a second reference to the class *Constraint*. Consequently UML could represent *StatementTemplate* as an association class, whilst the RDF language may define *StatementTemplate* as an auxiliary node. In the DSP model, *StatementTemplate* is represented as an auxiliary node that does not signify a named resource, i.e. a blank node or anonymous resource. In addition, following the same reasoning, *StatementTemplate* contains a reference to the class *Constraint*. *Constraint* is a nested structure that contains a collection of constraints and is also represented as an anonymous resource.

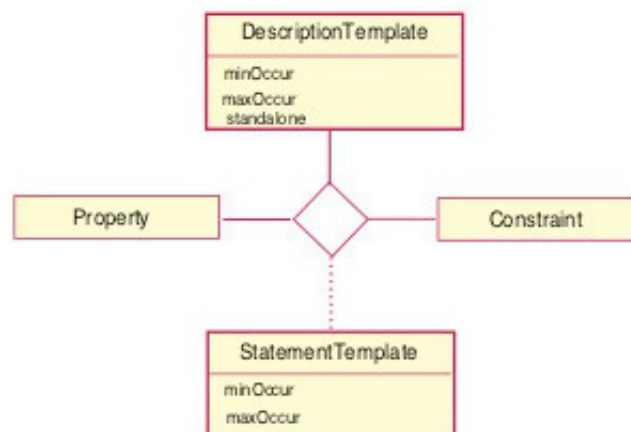


FIG. 1: UML class diagram describing the association class *StatementTemplate*

Blank nodes come with a significant overhead and additionally add unnecessary complexity to the DSP model.

As UML leverages the power of modeling effectively, we propose an extra approach to separate the modeling of the DSP into a three level hierarchy: a model at level 1 (Earth observation domain model), a meta-model at level 2 (DSP meta-representation) and a meta-metamodel at level 3 (RDF and RDFS meta-representations). We defend the claim that UML profiling mechanisms could help increase the usefulness of a RDF application profile particularly

in a linked open data context. An UML profile [(D'Souza, 1999)] represents a lightweight extension mechanism to the UML language by defining custom stereotypes in particular. Stereotypes are applied to UML elements to refine their semantics, either as classes or associations.

On that point, we take advantage of the work carried out on ontology metamodeling (Brockmans, 2006) with a corresponding UML profile and a collection of stereotypes that convey the meaning of the semantic web languages primitives (RDF, RDFS and OWL). Some of these stereotypes are illustrated in the simplified diagram of the DSP model depicted in FIG. 2.

A meta-class, as an example *RDFSClass*, *BlankNode* or *RDFProperty*, refines the semantics of each class of the DSP model. For instance, the generic class *BlankNode* marks appropriate dependencies on the classes *StatementTemplate* and *Constraint* that result from the translation of n-ary properties, respectively.

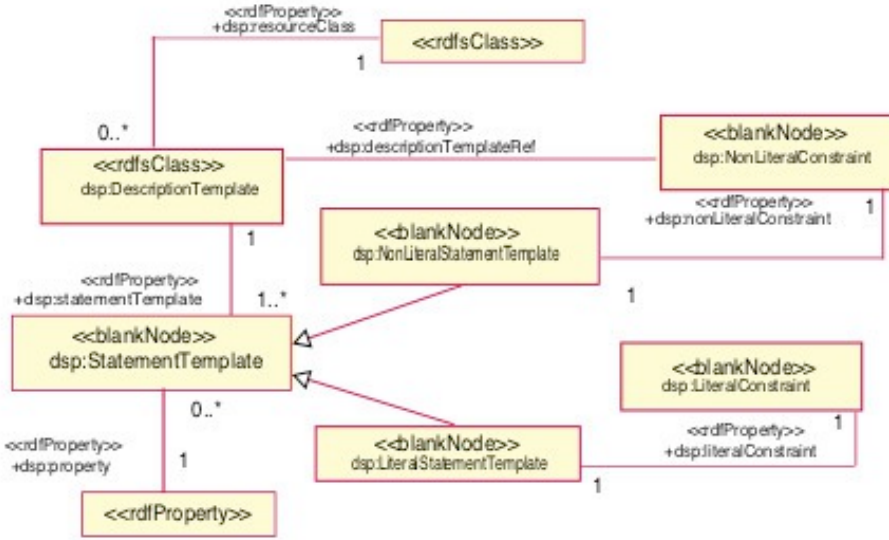


FIG. 2: Simplified DSP diagram qualified with the RDF/RDFS UML profile

3. Towards a RDF-compliant DSP model for Earth observations

We consider the constraint language DSP as a metamodel, associated with a dedicated UML profile that allows expressing the DSP semantic specifications by means of stereotypes. These stereotypes specialize the stereotypes described within the RDF/RDFS UML profile and give a specific way of defining typical constraints between DSP languages elements.

The main advantage of such an approach is to significantly raise the level of abstraction by providing nested models and different possibilities of zooms on the same elements of interest accordingly. We can therefore build a DSP model tailored to our specific domain by instantiating the DSP profile. The interest is two fold: we just have to concentrate on the description of the Earth observation resources and the approach is generic and may be reused for other thematic domains. FIG. 3 gives an excerpt of the DSP instantiation model for EO resources. The classes *EarthImage_T* and *Temporal_Extent_T* are annotated by the meta-class *dsp:DescriptionTemplate* that is a kind of *RDFSClass*. *EarthImage_T* refers to the class *EarthImage* through an association annotated with the meta-property *dsp:resourceClass*. Additionally *EarthImage_T* is supported by a number of unnamed classes marked by the meta-class *dsp:NonLiteralStatementTemplate*. One of these classes links the Earth Image description template to the second description template entitled *Temporal_Extent_T* and entirely dedicated to the temporal aspects. An object *Temporal_Extent_T* is connected through an object typed by *dsp:StatementTemplate* to an object

iso19108:Instant which corresponds to the date when the image is collected. The temporal metadata standards, in our case ISO 19108 are involved and will greatly facilitate the use of images in the context of multitemporal studies (see subsection 4.2). A similar outcome was achieved related to temporal aspects.



FIG. 3: Short excerpt of the EOAP model as a DSP profile instantiation

The same excerpt of the EOAP model is presented in Listing 1 on the following page using N3 syntax. The EOAP model pays particular attention to carefully describing the required level of temporal and spatial coverage in relation to real needs for Earth observing and uses relevant metadata standards. We illustrate in the section 4, the importance of the spatial and temporal dimensions by portraying various use cases of interest.

The contribution of the EOAP model together with the provision of multiresolution and multitemporal satellite imagery eases identification of environmental patterns over time and space.

4. Potential implementation of DCAP for image discovery and consultation purposes

4.1. Targeted users

The GEOSUD SDI is intended for French public stakeholders, whose missions contribute to environmental monitoring of French territories. The term “public stakeholder” covers a wide range of actors: those of research and higher education (research laboratory, university, high school) conducting studies, for instance, on the structure, functioning or dynamics of ecosystems.

Others actors include such decentralized or central governmental units, whose missions require, for instance the building of cartographic products to meet the monitoring or control of the implementation of governmental policies. Finally, local authorities are also public stakeholders whose missions have been recently extended to the management of the environment (waste management, biodiversity, air quality...) and for which high-resolution satellite imagery provides reference spatial datasets among others.

```

eoap:TemporalExtent_T
  a      dsp:DescriptionTemplate ;
  dsp:maxOccur      "infinity"^^xsd:nonNegativeInteger ;
  dsp:minOccur      "0"^^xsd:nonNegativeInteger ;
  dsp:resourceClass  iso19108:Period ;
  dsp:standalone    false ;
  dsp:statementTemplate [ a      dsp:NonLiteralStatementTemplate ;
    dsp:maxOccur      "1"^^xsd:nonNegativeInteger ;
    dsp:minOccur      "1"^^xsd:nonNegativeInteger ;
    dsp:nonliteralConstraint [ a      dsp:NonLiteralConstraint ;
      dsp:valueURIOccurrence "mandatory"^^dsp:occurrence ;
      dsp:vocabularyEncodingSchemeOccurrence
        "mandatory"^^dsp:occurrence ;
      dsp:vocabularyEncodingSchemeURI  iso19108:Instant ;
      dsp:property  iso19108:begin
    ] .
  ] .

eoap:EarthImage_T
  a      dsp:DescriptionTemplate ;
  dsp:maxOccur      "1"^^xsd:nonNegativeInteger ;
  dsp:minOccur      "1"^^xsd:nonNegativeInteger ;
  dsp:resourceClass  eoap:EarthImage ;
  dsp:standalone    true ;
  dsp:statementTemplate [ a      dsp:NonLiteralStatementTemplate ;
    dsp:maxOccur      "1"^^xsd:nonNegativeInteger ;
    dsp:minOccur      "0"^^xsd:nonNegativeInteger ;
    dsp:nonliteralConstraint [ a      dsp:NonLiteralConstraint ;
      dsp:descriptionTemplateRef  eoap:TemporalExtent_T ;
      dsp:valueURIOccurrence      "mandatory"^^dsp:occurrence ;
      dsp:vocabularyEncodingSchemeOccurrence
        "disallowed"^^dsp:occurrence ;
      dsp:property  dcterms:temporal
    ] .
  ] .

```

Listing 1: EOAP RDF Excerpt in N3

The variety of actors and the missions assigned to them point out the diversity of expertise and point of view, which our distribution platform must meet for image discovery and consultation purposes. The majority of these actors have little to no skills in remote sensing. Logically, their discovery and consultation requests will in the first stage be based on spatial and temporal properties of the satellite image. Others who have strong skills are able to have an expert approach and evaluate the appropriateness of an image from its characteristics (pixel resolution, format, encoding), or even those of the sensor or the conditions of acquisition of the image (incidence angle, cloud cover...)

4.2. Planned use cases

As described above, the target users are from very different domains and skills. The intended uses are equally variable, as they are designed to meet a wide range of environmental issues (see previous section). Among all these uses and to illustrate the relevance of our application profile in the GEOSUD SDI, we have chosen to describe three of them. The first two cases are about regulatory control missions for the management of renewable resources, the monitoring of land use planning by local governmental unit. The third case is an experiment from the scientific community with the aim to identify wetlands in tropical area.

4.2.1 Non expert use case: mapping and temporal monitoring of clearcuts² in Landes massif in the Southwest of France by decentralized governmental unit (DRAAF³)

In line with their missions and to enforce the regulations on exploitation of forest resources, DRAAF and DDT⁴ must implement control and monitoring tools, firstly, to establish control plans of clearcuts and secondly to ensure sustainable management of the forest resource. The production of a map identifying at time t the clearcuts of a forest is based on a methodology defined by (Ose, 2015). First, the mapping of clearcuts requires having two sets of vector data, one used to restrict the study area and the other to take into account the land use. The determination of clearcuts is based on two high-resolution satellite images acquired in the same season (preferably in spring) between an interval of two years or more. This is to calculate the difference of NDVI (Normalized Difference Vegetation Index) during the given interval and quantify the evolution of clear cuts. Taking the example of the DRAAF, who wants to establish the mapping in the Landes forest during the last two years, the GIS engineer knowing the clearcuts mapping methodology (but not a specialist in remote sensing) will want to discover the required images for his study with these words: "I am searching for high-resolution images that were acquired during the period from April to May for the years 2014 and 2015, covering the area from latitudes 43,97°; 45,06°- longitudes -1,56° ; -0,133°".

4.2.2 Non expert use case: mapping of artificial sprawl in the peri-urban zone of Montpellier by the national observatory of agricultural space consuming (ONCEA)

In France, urban sprawl dynamics are particularly strong. The increase and spreading of built-up areas towards the periphery takes place to the detriment of natural and agricultural spaces. The conversion of land with agricultural potential is cause of serious concerns as it is usually irreversible. Thus, for the land use planning services, the mapping of artificial sprawl dynamics is an essential tool for the quantification of lost agricultural space.

Based on the method of Dupuy et al. (2012), the monitoring of artificial sprawl is based on a very high resolution image provided by satellite such as Pleiades or SPOT6. It is also necessary to use French large-scale data repositories that provide both the state of the human impact of an area (roads, buildings) and the land use (forest, crops...). An object-oriented analysis of very high resolution image ensures the recognition of new buildings and transport infrastructure elements and thus quantifies the agricultural areas that were urbanized. The

ONCEA GIS engineer wants to build the mapping of artificial sprawl for the cities of Lattes and Pérols (southern of the city of Montpellier) to assess the evolution since 2012. To provide this, the engineer must have very high resolution images covering the area in question. These images should have a sub-metric resolution and must be acquired during the spring or summer. The object-oriented analysis is more efficient for this period. Also, we could formulate his request in our discovery application as: "I am searching for images with sub-metric resolution, acquired during the period from March to September, covering the cities of Lattes and Pérols".

The use of toponyms to select the images overlapping the study area takes advantage of semantic external resources. In this case, we use the places ontology called Geonames (<http://sws.geonames.org/>). It allows us to match the spatial footprint of an image, expressed in geographic coordinates, with those of cities supplied by Geonames ontology. Finally, we can annotate the images with the names of the cities that are included or that intersect the spatial extent of an image.

² Clearcut: refers to a mode of forestry development through cutting down of all trees of a parcel

³ DRAAF: Regional Headquarter for food, agriculture and forest

⁴ DDT: Sub-regional headquarter for territory management

4.2.3 Expert use case: Discrimination of wetlands in Madagascan forest by a remote sensing specialist

Proposed by [Hajalalaina, 2013] the identification of wetlands in the Madagascan forest meets agronomic and environmental issues. The wetlands are potential areas for rice crops and also biodiversity reserve. The proposed method is based on the use of multi-source and multi-resolution images. The LandSat7ETM+ product, at 30 meters of resolution, will as a first step allow drawing up a map of wetlands at regional level through a classification of the image pixel. In a second step, an object-oriented classification is applied to a high resolution image (2.5 meters), namely SPOT5 image. This second step results in a mapping of wetlands at the local level. In order to have these two kinds of images available, the remote sensing specialist will formulate his query specifying the name of the sensors required which provide the expected spatial resolution for the determination of wetlands. The specialist will also build his request on the characteristics of the image in order to define the temporal and spatial extents over which he wishes to conduct his study. Thus, we could formulate the entire request as: "I am searching for images acquired by Landsat 7 platform and the images acquired by the SPOT5 in panchromatic mode whose spatial footprints are between the latitudes -20,58° ; -22,35° and longitudes 47,85° ; 46,44°, which were acquired between the month May and June."

5. Conclusions and future work

The work carried out results in the definition of a Dublin Core application profile that is intended to meet the needs of different actors with regard to both satellite imagery uses and environmental issues. The use cases that we have examined reveal high requirements in capabilities to access, query and analyze a significant number of series of satellite images. The application profile EOAP is hence logically supported by metadata standards that are specifically dedicated either to spatial and temporal dimensions or to descriptions of observations and measures. An image is above all a digital resource and EOAP is also drawing on Dublin Core Metadata Element Set.

Additionally we initiate work towards meta-modeling activities to complete the RDF-based DSP model with higher levels of abstraction to efficiently drive the building of a thematic model that conforms to the DSP model. We will continue our modeling efforts focusing on two main directions. First UML profiling activities could constitute an efficient way to design an application before committing to implementation. We will therefore develop a generic RDF-based editor to build DSP models from the defined UML profiles. Secondly, we will add some constraints based on OCL (Object Constraint Language) (Clark, 2002) to the DSP metamodel. These constraints will be used for the validation of instantiations.

Acknowledgements

This work was supported by public funds received in the framework of GEOSUD, a project (ANR-10-EQPX-20) of the program "Investissements d'Avenir" managed by the French National Research Agency.

References

- Ahlers, Dirk, 2013. Assessment of the accuracy of GeoNames gazetteer data. In Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13), Chris Jones and Ross Purves (Eds.). ACM, New York, NY, USA, 74-81
- Brockmans, Saartje, Robert M. Colomb, Peter Haase, Elisa F. Kendall, Evan K. Wallace, Chris Welty, and GuoTong Xie. (2006). A Model Driven Approach for Building OWL DL and OWL Full Ontologies. In The Semantic Web- ISWC 2006 (pp. 187-200). Springer Berlin Heidelberg, 2006.

- Clark, Tony, and Jos Warmer, editors. (2002). Object Modeling with the OCL: The Rationale behind the Object Constraint Language. Vol. 2263 in LNCS. Springer-Verlag, 2002.
- CWA 15248 Guidelines for machine-processable representation of Dublin Core Application Profiles, April 2005 DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved April 10, 2015, from <http://www.dublincore.org/documents/1998/09/dces/>.
- Desconnets, J. Christophe, Hatim Chahdi, and Isabelle Mougenot. (2014). Application Profile for Earth Observation Images. Metadata and Semantics Research, (pp. 68-82). Springer International Publishing.
- D'Souza, Desmond, Aamod Sane, and Alan Birchenough. (1999). First-class extensibility for UML - packaging of profiles, stereotypes, patterns. In Robert France and Bernhard Rumpe, editors, UML'99 - The Unified Modeling Language, volume 1723 of Lecture Notes in Computer Science, (pp. 265-277). Springer Berlin Heidelberg, 1999.
- Dublin Core Metadata Element Set, Version 1.1. (2012). Retrieved April 10, 2015, from <http://www.dublincore.org/documents/dces/>.
- Dupuy, Stéphane, Eric Barbe, and Maud Balestrat. (2012). An object-based image analysis method for monitoring land conversion by artificial sprawl use of RapidEye and IRS data. Remote Sensing 4.2, 2012, 404-423.
- Fowler, Martin. 2010. Domain Specific Languages (1st ed.). Addison-Wesley Professional.
- Gasperi, Jerome, Frédéric Houbie, Andrew Woolf, and Steven Smolders. (2012). Earth observation metadata profile of observations & measurements. OGC Document Number 10-157r3 ,2012.
- Hayes, Patrick, and Brian McBride. (2004). RDF semantics. Retrieved April 10, 2015, from <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/>.
- Hajalalaina, Aimé Richard, Manuel Grizonnet, Eric Delaître, Solofo Rakotondraompiana, and Dominique Hervé. (2013). Discrimination des zones humides en forêt malgache, proposition d'une méthodologie multirésolution et multisource utilisant ORFEO toolbox. Revue Française de Photogrammétrie et de Télédétection, 2013,(pp. 37-48).
- ISO, (2003) : ISO 19115 - Geographic Information - Metadata, International Standard Organization. First edition, 2003-05-01
- ISO, (2009) : ISO 19115-2 - Geographic Information - Metadata part 2: extension for imagery and gridded data, International Standard Organization, 2009. First edition. 2009-02-15
- Kazmierski, Mathieu, J. Christophe Desconnets, Bertrand Guerrero, and Dominique Briand. GEOSUD SDI. (2014). Accessing Earth Observation data collections with semantic-based services. In Proceedings of the 17th AGILE Conference on Geographic Information Science, Connecting a Digital Europe through Location and Place, Castellon, Spain, June 2014.
- Nilsson, Mickael, Alistair J. Miles, Pete Johnston, and Fredrik Enoksson. (2009). Formalizing Dublin Core Application Profiles-Description Set Profiles and Graph Constraints. In Metadata and Semantics (pp. 101-111). Springer US.
- Ose, Kenji, and Michel Deshayes. (2015). Détection et cartographie des coupes rases par télédétection satellitaire. Guide méthodologique. Version 5.0. UMR Tetis, IRSTEA. 2015.
- Rumbaugh, James, Michael Blaha, William Premerlani, Frederick Eddy, and William Lorensen. (1991). Object-oriented modeling and design. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1991.
- Warren, Rob and Erik Champion. (2014). Linked Open Data Driven Game Generation. In Proceedings of the 13th International Semantic Web Conference - Part II (ISWC '14), Springer-Verlag New York, Inc., New York, NY, USA, (pp. 358-373), 2014.

A DCAP for the Social and Solidarity Economy

Mariana Curado Malta
ISCAP/CEISE, Portugal
Algoritmi Center,
Portugal
mariana@iscap.ipp.pt

Ana Alice Baptista
Algoritmi Center,
Portugal
analice@dsi.uminho.pt

Cristina Parente
Instituto de Sociologia,
Universidade do Porto,
Portugal
cparente@letras.up.pt

Abstract

This article presents a work-in-progress version of a Dublin Core Application Profile (DCAP) developed to serve the Social and Solidarity Economy (SSE). Studies revealed that this community is interested in implementing both internal interoperability between their Web platforms to build a global SSE e-marketplace, and external interoperability among their Web platforms and external ones. The Dublin Core Application Profile for Social and Solidarity Economy (DCAP-SSE) serves this purpose. SSE organisations are submerged in the market economy but they have specificities not taken into account in this economy. The DCAP-SSE integrates terms from well-known metadata schemas, Resource Description Framework (RDF) vocabularies or ontologies, in order to enhance interoperability and take advantage of the benefits of the Linked Open Data ecosystem. It also integrates terms from the new essglobal RDF vocabulary which was created with the goal to respond to the SSE-specific needs. The DCAP-SSE also integrates five new Vocabulary Encoding Schemes to be used with DCAP-SSE properties. The DCAP development was based on a method for the development of application profiles (Me4MAP). We believe that this article has an educational value since it presents the idea that it is important to base DCAP developments on a method. This article shows the main results of applying such a method.

Keywords: Application Profile; interoperability; Metadata schemas, Vocabulary Encoding schemes, Social and Solidarity Economy.

1. Introduction

This article presents a work-in-progress Dublin Core Application Profile developed to serve the Social and Solidarity Economy (SSE) sector –DCAP-SSE V1.1– referred to hereafter as DCAP-SSE. Cooperatives, associations and mutualities, among others, are types of organizations that belong to this sector. The SSE is different from the economy of State and Market (Lechat, 2007) since it is created by an organised civil society. SSE organisations are interested in developing activities for the common good, with the goals of SSE organisations being neither centered in profit nor in individualistic needs. Therefore, SSE presents itself as a material and human alternative to capitalist economy (Cattani, Laville, Gaiger, & Hespanha, 2009). SSE, according to the spatio and temporal contexts, can take on other names such as the “third sector” used for example in the USA and Europe, or “non-governmental organisations” (NGO) widely used in the field of aid for development in peripheral countries.

SSE organisations work with scarce resources, therefore networking and partnerships appear as a highly relevant way of working, allowing SSE organisations to gain visibility and attract funding, or even to be able to work at scale.

These organisations have machine-to-machine communication needs that are internal or external to them, for example, to other kinds of organisations such as governmental agencies. In order to support these machine-to-machine communication needs, there is the need to provide interoperable solutions among the software platforms that support their activities. There are several approaches to interoperability. In the context of information technologies, interoperability

can be defined as the possibility of multiple systems, with different kinds of software or hardware, and different data structures and interfaces, to exchange data without previous communication, with the minimum loss of contents and functionality (NISO, 2004, p.1). The Dublin Core Metadata Initiative (DCMI) defines interoperability in its glossary as: “The ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner. There are three aspects of interoperability: semantic, structural and syntactical” (DCMI, 2011). For more information about interoperability see, for example, Institute of Electrical and Electronics Engineers (2010); Interoperable Delivery of European eGovernment Services to public Administrations Businesses and Citizens (2004); Payette, Blanchi, Lagoze, & Overly (1999).

Semantic interoperability focuses on meaningful exchanges of information, i.e., information that has the same interpretation (or very closely) by both the sender and the receiving systems. Our work is carried out under this perspective and in the context of the Semantic Web.

The Semantic Web has technologies that “enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data is empowered by technologies” that started to emerge in 1999. It is about common formats for integration and combination of data from different sources (W3C, 2012). This data is mostly what is being called metadata, in the way that it is “data about data” (DCMI, 2011) and follows well-defined rules of metadata schemas. A metadata schema is a set of “metadata elements designed for a specific purpose, such as describing a particular type of information resource” (NISO, 2004, pp. 4).

In order to provide “a foundation for the development of application-independent syntax specifications and constraint languages”, DCMI developed the Dublin Core Abstract Model (DCAM) (Powell, Nilsson, Naeve, Baker, & Johnston, 2007) that presents the components and constructs used in DCMI metadata. One of these constructs is the Dublin Core Application Profile (DCAP) - “a generic construct for designing metadata records” (Baker & Coyle, 2009), a DCAP describes “the structure and contents of data” (Baker & Coyle, 2013). The definition of rules to build a DCAP is set in the Singapore Framework for Dublin Core Application Profiles, a DCMI Recommendation (c.f. Nilsson, Baker, & Johnston (2008)). This DCMI work has been developed under the umbrella of international standards. The use of these international standards is critical when it comes to semantic interoperability, but it is not sufficient, since a community needs to follow some rules to achieve high levels of interoperability. These rules are defined in the interoperability layers model (c.f. Nilsson, Baker, & Johnston (2009)), which allows a community to assess the “interoperability reach” of a particular implementation. The interoperability layers model defines 4 levels of interoperability that have to do with the use of: (i) metadata schemas and DCMI vocabularies, in levels 1 and 2; and (ii) DCMI standards: DCAM and DCAP, in levels 3 and 4. Level 4 is the highest level of interoperability which is achieved when a community uses the DCAP construct as a reference and binding to describe its resources. Thus, a DCAP became a very important instrument to implement interoperability.

A recent study by Curado Malta, Baptista, & Parente (2014) reveals that the SSE community is facing a global challenge. This community wants to implement interoperability between their Web platforms –to build a global SSE e-marketplace– and also among their Web platforms and external ones. After a study of the environment, its requirements, and its internal and external constraints, we came to the conclusion that there was no DCAP that could serve the SSE community. SSE organisations are submerged in the market economy but they have specificities that were not taken into account in the market economy but which are very important for SSE.

At the end of 2010 the Intercontinental Network for the promotion of Social and Solidarity Economy (RIPESS)¹ created a task force called ESSGlobal for the development of interoperability among its members’ platforms, and decided to develop a DCAP.

¹ RIPESS. <http://www.ripest.org> (accessed on January 20, 2015).

The DCAP-SSE integrates not only terms from well-known metadata schemas, Resource Description Framework (RDF) vocabularies or ontologies (namely dcterms, foaf, vcard, schema.org and good relations), in order to enhance interoperability and take advantage of the benefits of Linked Open Data ecosystem (LOD), but also terms from the essglobal RDF vocabulary. This new vocabulary was created to respond exactly to the SSE specific needs—e.g., a pre-requisite of the SSE is the open cost which is a breakdown of all inputs, such as taxes and raw materials, and labour costs that make up the product or service's final cost. The DCAP-SSE also integrates five new Vocabulary Encoding Schemes created by the DCAP-SSE development-group to be used with DCAP-SSE properties.

This article proceeds as follows: Section 2 presents the methodology used to develop the DCAP-SSE; section 3 presents the DCAP-SSE: the functional requirements, the domain model and the Description Set Profile, and some other technical information we consider relevant. The last section presents conclusions and future work.

2. Methodology

A DCAP development can be a complex task since it happens in a completely open environment. In addition to that, this kind of development is often framed in multi-cultural-organizational-language environments. This work is no exception. In fact, the ESSGlobal development-team integrates persons with different profiles: 7 SSE experts, 3 data modelers and 1 Semantic Web expert: the SSE experts belong to different organisations of the RIPESS network with top organisations of SSE in Brazil, Canada, France, Italy, Luxemburg, Spain and USA. Two data modelers were members of EITA², a Brazilian SSE cooperative; the researcher leading the development-team was from the Algoritmi Research Center in Portugal,³ a data modeler and the Semantic Web expert.

The SSE organisations participating in the DCAP-SSE development differ in organization-type, location, culture and in the language they speak. To find a common ground of understanding in such an environment becomes a huge challenge. We think that the existence of methods for the development of a DCAP may help to address this challenge. The DCAP-SSE development work was framed in a PhD research project (Curado Malta & Baptista, 2013a, 2013b; Curado Malta, 2014) that resulted in the definition of a method for the development of metadata application profiles (Me4MAP). This project was based on a design science research methodology, with the framework defined by Hevner & Chatterjee (2010). The DCAP-SSE development work was the experimental situation defined by Hevner & Chatterjee (2010) to test the artifact in development (Me4MAP). The development of DCAP-SSE was informed by the development of Me4MAP and vice versa. The focus of this article is the DCAP-SSE. A fuller explanation of Me4MAP is in preparation.

According to Me4MAP, a DCAP development should follow the Singapore Stages. The name of the stages are based in the seminal document *The Singapore framework for Dublin Core Application Profiles* (c.f. Nilsson et al. (2008)). This framework defines three mandatory Singapore Components: Functional Requirements; Domain Model and Description Set Profile, and two optional components: Usage Guidelines and Syntax Guidelines. This framework does not define a sequence of activities, but in fact the Singapore Components have a logic order of development and every Component builds upon the previous one. A method organizes the activities in a sequence and Me4MAP does the same. Unlike other methods, each activity results in a deliverable which are the Singapore Components already referred.

The DCAP-SSE development was carried out as follows:

- In the first Stage we developed the Functional Requirements. This activity included the sub-activities of: (i) definition of the *vision* of the project; (ii) definition of the

² EITA. <http://www.eita.org.br> (accessed July 2, 2015).

³ Algoritmi Research Center. <http://algoritmi.uminho.pt> (accessed July 2, 2015).

application domain; (iii) elicitation of the *high-level requirements*; (iv) development of the *use-case model*, and (v) the elicitation of the *functional requirements*.

- In the second Stage we developed the *domain model*. This activity included the sub-activities (i) definition of the *environmental scan*, and (ii) definition of the *domain model*;
- In the third Stage, we developed the Description Set. This activity included the sub-activities of:
 - i. development of Pre-Description Set profile including sub-activities of defining the:
 - a) Detailed Domain Model;
 - b) Vocabulary Alignment; and
 - c) Constraints Matrix;
 - ii. encoding of the Description Set Profile.

The next section shows the DCAP-SSE Singapore Components and some of the deliverables that led to the definition of these Components.

3. Dublin Core Application Profile for the Social and Solidarity Economy

The DCAP-SSE development project's wiki page⁴ includes DCAP-SSE's technical information.

As in any other projects, it is very important to set boundaries in order to effectively identify the issues the project aims to address. To accomplish this task, the DCAP-SSE team defined a Vision Statement as follows:

“ESSglobal is an initiative of some RIPESS members with the following objectives:

- Increase the international visibility of the activities and products of solidarity economy;
- Pool the methods and tools of mapping projects that already exist and that are being developed;
- Develop transversal projects of human and economic cooperation among the participants of the working group;
- Cooperate with other initiatives (existing or being created) that specialize in information systems, in the geo-referencing of SSE actors, and in networking.

The DCAP-SSE covers the following dimensions:

- Commerce;
- Public visibility;
- Research and statistics;
- Network building;
- Public policies;
- Education.

The dimensions of “Education” and “Public policies” are not present in this first version of the DCAP-SSE.”

3.1. Functional Requirements

As already mentioned in the methodology section, the Me4MAP suggests that the Functional Requirements should be developed based on all activities of the first Stage, especially on the identified uses cases.

The Functional requirements defined for DCAP-SSE are to:

⁴ Project Wiki. <http://purl.org/essglobal/wiki/> (accessed on April 4, 2015).

- enable the creation and sharing of consistent metadata;
- support the search by any or all items: “SSEInitiative”, “Network”, “Product”, “Sale Options” and “Product-Input”: these functional requirements meet the needs of Use Cases 1, 2 and 3 which are described in the Project Wiki;⁵
- support the search for any property of each element mentioned in the previous paragraph and also “Cost Composition” of any Product-Input: these functional requirements meet the needs of the previously referenced Use Cases 1, 2 and 3 on the Project Wiki.

3.2. Domain Model

Figure 1 presents DCAP-SSE domain model as an Object Role Model (ORM) diagram.

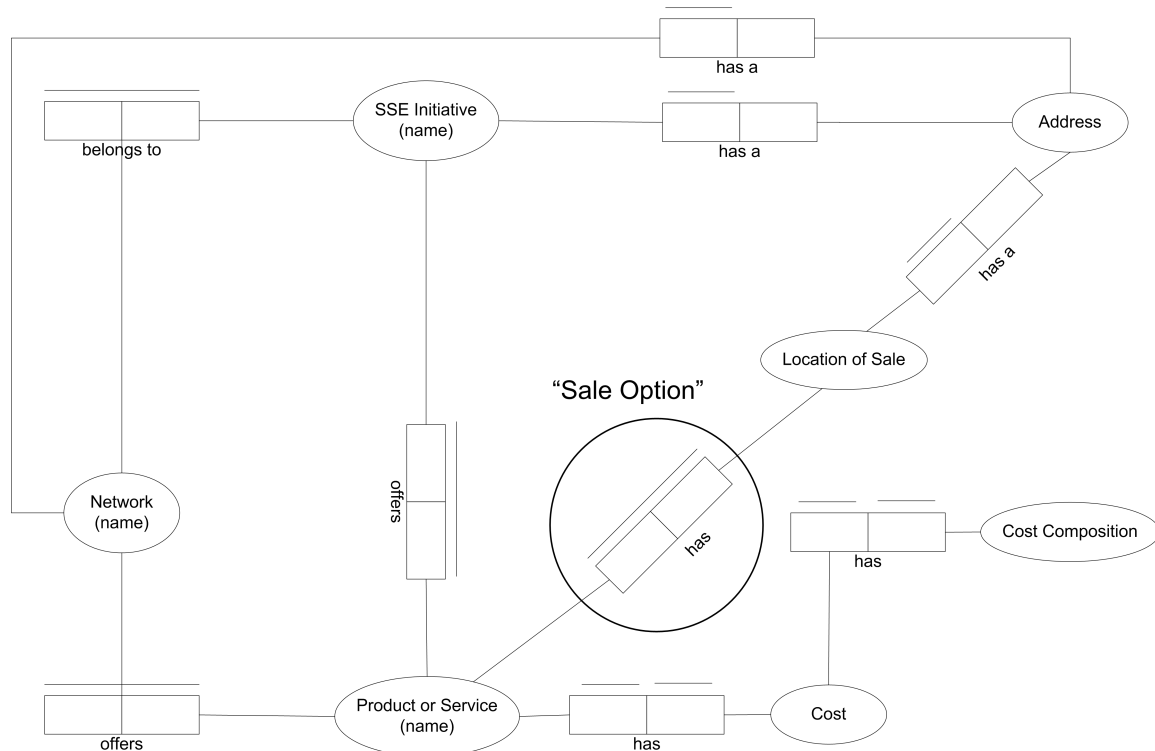


FIG 1. DCAP-SSE Domain Model

This domain model represents eight entities: SSE Initiative, Network, Product or Service, Cost, Cost Composition, Location of Sale, Sale Option and Address; and the relations between them. The next section provides more details about the classes and relations.

3.3. Description Set Profile

The Description Set Profile (DSP) of DCAP-SSE can be accessed online.⁶ The DCAP-SSE integrates terms from well-known metadata schemas, Resource Description Framework (RDF) vocabularies or ontologies in order to enhance interoperability and take advantage of the benefits of Linked Open Data (LOD) ecosystem - see Table 1.

Figure 2 in appendices to this article, presents the Unified Modeling Language (UML) diagram showing the data model of the relations between the DCAP-SSE terms. This model shows the information we want to gather, the definition of: (i) the organisations and its networks; (ii) the products or services sold by the organisations or by the networks; (iii) the several components of the open cost of the products or services; and (iv) the location of sale of the products or services.

⁵ Wiki: Use Cases. <http://www.maltas.org/wiki-essglobal/doku.php?id=use>.

⁶ DSP. <http://purl.org/essglobal/dsp-xml> (accessed on 04 April, 2015).

TABLE 1: Metadata schemas used in the DCAP-SSE

Title	Namespace URI	Prefix
DC TERMS	http://purl.org/dc/terms/	dcterms
The friend of a friend	http://xmlns.com/foaf/0.1/	foaf
Good Relations	http://purl.org/goodrelations/v1#	gr
VCARD	http://www.w3.org/2006/vcard/ns/	v
Schema.org	http://schema.org	schema
Social and Solidarity Economy	http://purl.org/essglobal/vocab/	essglobal

DCAP-SSE has the following classes:

- **SSEInitiative**: an organization, practice, network, or other initiative that is recognized as belonging within the SSE;
- **Network**: a network of individuals and/or organisations that participate in the SSE;
- **ProductOrService**: the good offered by an **SSEInitiative** or **Network**. It may be material good or provision of service;
- **LocationOfSale**: a place where the goods or services of an SSE initiative are provided. It can be self-owned shops, but also SSE partner places where products or services are available among those from other initiatives;
- **Address**: the physical address of a **LocationOfSale**, of a **SSEInitiative** or of a **Network**;
- **SaleOption**: a product or service sold at a given price, under specific properties, in a given **LocationOfSale**. The delivery costs are included in this class;
- **Cost**: the final cost for a particular product or service produced by an **SSEInitiative** or **network**, including all costs components. The price will be this cost added to delivery costs and sales margin;
- **CostComposition**: a breakdown of all inputs (such as taxes and raw materials) and labour costs that make up the product or service's final cost;
- **Input**: a product, service, or activity that goes into making the final product or service;
- **Labour**: work done for specific tasks related to the provision of goods or services offered by the **SSEInitiative**. Generally it can be human, animal or machine labour, but **ESSglobal** considers human labour only;
- **OtherCosts**: other costs which impact on the final cost of a product or service provided by an **SSEInitiative** other than **Input** or **Labour**, like taxes, depreciation of machinery, funds, etc.

Details about the: i) properties related to each class, and the properties that relate classes; ii) cardinality of each property; and iii) constraints of each property can be found in the Constraints Matrix deliverable in the Project Wiki.⁷ This matrix is based on the table presented in the Guidelines for DCAP by Baker & Coyle (2009) with some adjustments and improvements. An excerpt of this matrix is presented in Figure. 3, in appendices to this article.

3.4. ESSGlobal RDF Vocabulary and Vocabulary Encoding Schemes created

The ESSGlobal development-team did not find terms in the metadata community that could describe some of the SSE community specificities. SSE organisations, despite being submerged in the market economy, need to describe their resources taking into account dimensions such as: (i) the description of specific characteristics of the SSE organisations; (ii) the description of relations and networks that exist among SSE organisations; (iii) the description of the product or service's open cost, i.e. the breakdown of all inputs (such as taxes and raw materials) and labour

⁷ Constraints Matrix. <http://purl.org/essglobal/wiki/> (accessed on April 4, 2015).

costs that make up the product's or the service's final cost. In SSE, these costs are included and differentiated (as open cost) in the final price of the products or services.

The essglobal RDF vocabulary was created in order to fill these gaps. This vocabulary is available online⁸, it was registered in the Linked Open Vocabularies (LOV) platform.⁹

The essglobal RDF vocabulary has:

- 11 classes: 4 of these classes are sub-classes of well-known RDF vocabularies classes;
- 29 properties: 9 are object properties and the remaining 20, datatype properties.

Five vocabulary encoding schemes (VES) were created to be used with some of the DCAP-SSE properties (links accessed on 04 April, 2015):

- Economic Activities/Sectors¹⁰
- Macro-themes¹¹
- Qualifiers¹²
- Type of Labour¹³
- Legal form¹⁴

4. Conclusions and future work

The DCAP-SSE explicated here was developed based on a method for the development of DCAP (Me4MAP). A fuller explanation of Me4MAP is in preparation to be published in the future. We believe that this article has an educational value since it presents the idea that it is important to base DCAP developments on a method and shows the main results of applying such a method.

We think that the primarily use of SSE metadata will be to aid the discovery of SSE goods or services and networks, and for calculating statistical data (e.g. types of organisations, gender distributions of workers, etc). We predict that, in the first years of deployment, this data will be mostly about describing organisations (numbers employed, objectives, mission, address, membership in networks) and in a near future, it will also be about the goods or services offered. An example of application of the available SSE metadata could be the development of Apps for smartphones that can present users with the location and characteristics of nearby SSE organisations.

As future work, we will follow three tracks: a Research track, a Marketing & Technical support track and a Development Track:

- Research Track: the DCAP-SSE version presented in this article is a work-in-progress version since there are still steps to accomplish: i) a laboratory validation with samples from different SSE Web platforms; ii) a revision of DCAP-SSE after the laboratory validation; iii) inclusion of new dimensions, and new organisations in the development team, in order to enrich the DCAP-SSE expressivity.
- Marketing & Technical support Track: the DCAP-SSE development team is aware of the need to define and implement a dissemination plan for the SSE global community: there is the need to find ways to explain the potential of this new tool in a community that works with so few resources. On the other hand, SSE organisations that are willing to enter the LOD ecosystem will need technical support in order to understand how to use the DCAP-SSE. In order to achieve this we will need to develop manuals and use cases.

⁸ RDF vocabulary: <http://purl.org/essglobal/vocab/> (accessed on 04 April, 2015).

⁹ LOV: <http://lov.okfn.org/dataset/lov/vocabs/essglobal> (accessed on 04 April, 2015).

¹⁰ Economic Activities/Sectors: <http://purl.org/essglobal/standard/activities>

¹¹ Macro-themes: <http://purl.org/essglobal/standard/themes>.

¹² Qualifiers: <http://purl.org/essglobal/standard/qualifiers>.

¹³ Type of Labour: <http://purl.org/essglobal/standard/type-of-labour>.

¹⁴ Legal form: <http://purl.org/essglobal/standard/legal-form>.

- Development Track: there is the need to reflect on User Interface developments or ways to present the SSE metadata within an application framework for the SSE community.

Acknowledgements

Part of this work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013. We acknowledge the RIPESS – ESSGlobal task group for the work and knowledge they have shared with us. We acknowledge Bernard Vatant's help on the configuration of the ESSGlobal RDF vocabulary to be registered in the Linked Open Vocabularies (LOV) platform.

References

- Baker, T., & Coyle, K. (2009). Guidelines for Dublin Core Application Profiles. Retrieved from <http://dublincore.org/documents/profile-guidelines/>
- Baker, T., & Coyle, K. (2013). Application Profiles as an alternative to OWL Ontologies. Retrieved February 10, 2015, from http://wiki.dublincore.org/index.php/APvOWL_Lisbon
- Cattani, A. D., Laville, J.-L., Gaiger, L. I., & Hespanha, P. (2009). *Dicionário Internacional da Outra Economia*. CES.
- Curado Malta, M. (2014). *Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web Semântica*. University of Minho - Escola de Engenharia - Tecnologias e Sistemas de Informação. Retrieved from <http://hdl.handle.net/1822/30262>
- Curado Malta, M., & Baptista, A. A. (2013a). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In *International Conference on Dublin Core and Metadata Applications*. (pp. 90–103). DCMI. Retrieved from <http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81>
- Curado Malta, M., & Baptista, A. A. (2013b). Me4DCAP V0. 1: A method for the development of Dublin Core Application Profiles. *Information Services and Use*, 33(2), 161–171. doi:10.3233/ISU-130706
- Curado Malta, M., Baptista, A. A., & Parente, C. (2014). Social and Solidarity Economy Web Information Systems: State of the Art and an Interoperability Framework. *Journal of Electronic Commerce in Organizations*, 12, 35–52. doi:<http://dx.doi.org/10.4018/jeco.2014010103>
- DCMI. (2011). DCMI Glossary. Retrieved from <http://wiki.dublincore.org/index.php/Glossary>
- Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems - Theory and Practice*. (R. Sharda & S. Voß, Eds.) (Integrated). Springer.
- Institute of Electrical and Electronics Engineers. (2010). Standards glossary. Retrieved from http://www.ieee.org/education_careers/education/standards/standards_glossary.html
- Interoperable Delivery of European eGovernment Services to public Administrations Businesses and Citizens. (2004). Interoperability for european egovernment services. Retrieved from <http://ec.europa.eu/idabc/en/document/5313/5883.html>
- Lechat, M. P. (2007). Economia social, economia solidária, terceiro setor: do que se trata? *Civitas- Revista de Ciências Sociais*, 2(1), 123–140.
- Nilsson, M., Baker, T., & Johnston, P. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved February 10, 2015, from <http://dublincore.org/documents/singapore-framework/>
- Nilsson, M., Baker, T., & Johnston, P. (2009). Interoperability Levels for Dublin Core Metadata. Retrieved February 10, 2015, from <http://dublincore.org/documents/interoperability-levels/>
- NISO. (2004). *Understanding Metadata*. National Information Standards. Bethesda, USA: NISO Press.
- Payette, S., Blanchi, C., Lagoze, C., & Overly, E. A. (1999). Interoperability for digital objects and repositories. *D-Lib Magazine*, 5(5), 1082–9873.
- Powell, A., Nilsson, M., Naeve, A., Baker, T., & Johnston, P. (2007). DCMI Abstract Model. Retrieved from <http://dublincore.org/documents/2007/06/04/abstract-model/>
- W3C. (2012). W3C Semantic Web Activity. Retrieved February 10, 2015, from <http://www.w3.org/2001/sw/>

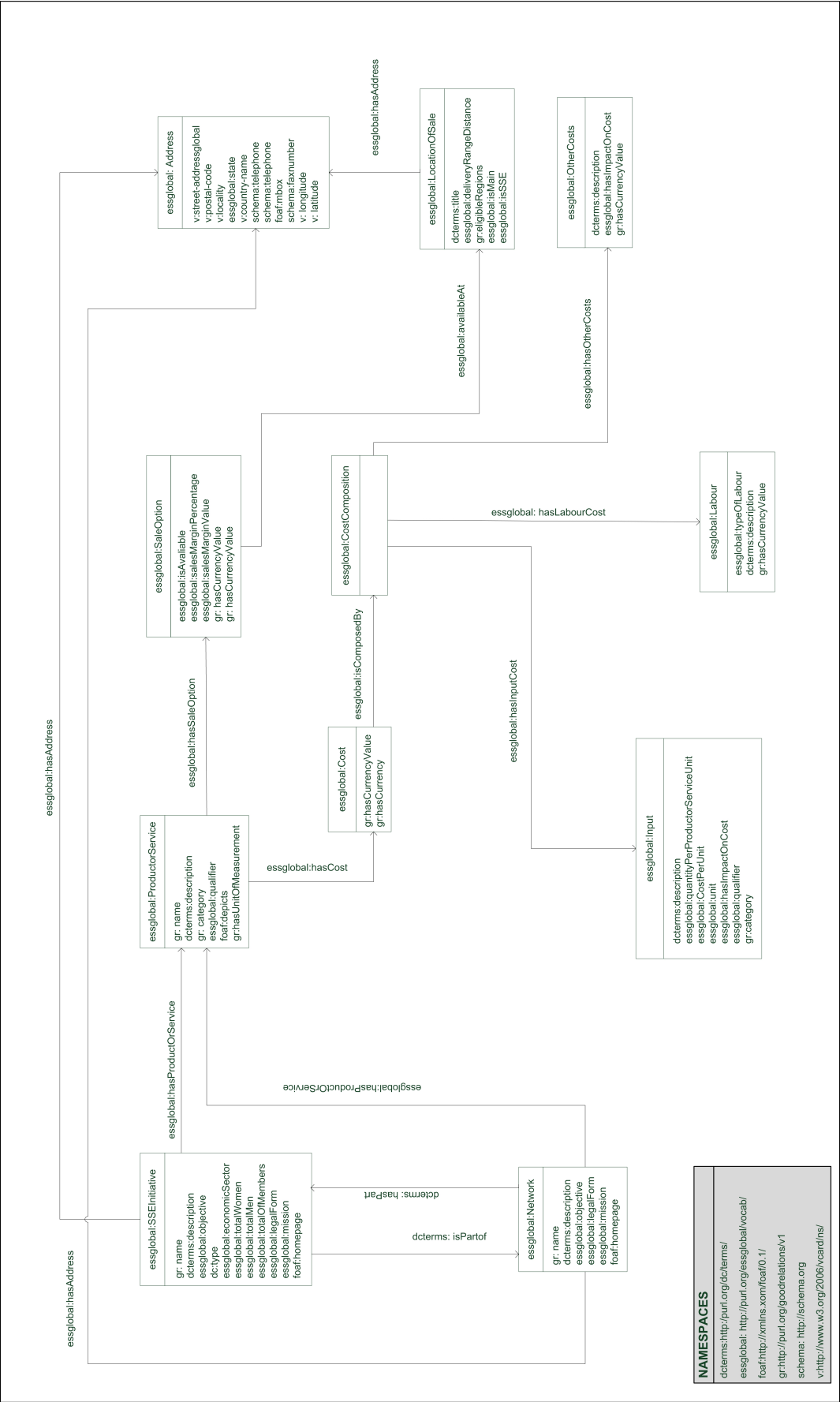


FIG 2. DCAP-SSE terms relations

Definition of Description Templates											
Description Template:	SSE initiative	Term	essglobal:SSEInitiative				Usage:	An organization, practice, network, or other initiative that is recognized as belonging within the social solidarity economy.			
Label	Term	Range	Value String	SES URI	Value URI	VES URI	Related description	Min	Max	Type	Usage
Name	gr:name	literal	YES	no	n/a	n/a	n/a	1	1	n/a	The name of the SSE initiative in its original language
Description	dc:terms:description	literal	YES	no	n/a	n/a	n/a	0	1	n/a	A description of the SSE initiative
Objective	essglobal:objective	literal	YES	no	n/a	n/a	n/a	0	1	n/a	Actionable and measurable steps the initiative is taking to carry out its mission.
Mission	essglobal:mission	literal	YES	no	n/a	n/a	n/a	0	1	n/a	The initiative's vision, values, and principles.
Type	dc:type	literal	YES	no	n/a	n/a	n/a	0	infinity	n/a	
Economic Sector	essglobal-economicSector	non-literal	no	no	YES	http://purl.org/essglobal/standard/activities	n/a	0	infinity	n/a	The economic sector of the SSE Initiative. ESSGlobal has established a VES form this term.
Total of Women	essglobal-totalWomen	literal	YES	no	n/a	n/a	n/a	0	1	n/a	The number of women that work in the SSE Initiative.
Total of Men	essglobal-totalMen	literal	YES	no	n/a	n/a	n/a	0	1	n/a	The number of men that work in the SSE Initiative.
Total of Members	essglobal-totalOfMembers	literal	YES	no	n/a	n/a	n/a	0	1	n/a	The total number of members that work in the SSE Initiative.
Legal Form	essglobal-legalForm	non-literal	no	no	YES	http://purl.org/essglobal/standard/legal-form	n/a	0	1	n/a	
URL	foaf:homepage	non-literal	YES	no	n/a	n/a	n/a	0	1	n/a	The initiative's homepage
Has Product or Service	essglobal-hasProductOrService	non-literal	YES	no	no	no	YES	0	infinity	n/a	The good offered by the SSE initiative.
has address	essglobal:hasAddress	non-literal	YES	no	no	no	YES	0	1	n/a	The address of the SSE Initiative
Is part of	dc:terms:isPartOf	non-literal	YES	no	no	no	YES	0	infinity	n/a	A network of individuals and/or organizations that the SSE Initiative belongs to.
Description Template:											
Label	Network	Term	essglobal:Network				Usage:	A network of individuals and/or organizations that participate in the SSE.			
Label	Term	Range	Value String	SES URI	Value URI	VES URI	Related description	Min	Max	Type	Usage
Name	gr:name	literal	YES	no	n/a	n/a	n/a	1	1	n/a	Name of the network
Description	dc:terms:description	literal	YES	no	n/a	n/a	n/a	0	1	n/a	A description of the network
Objective	essglobal:objective	literal	YES	no	n/a	n/a	n/a	0	1	n/a	Actionable and measurable steps the initiative is taking to carry out its mission.
Legal Form	essglobal-legalForm	non-literal	no	no	YES	http://purl.org/essglobal/standard/legal-form	n/a	0	1	n/a	ESSGlobal has established a VES for this property.
Mission	essglobal:mission	literal	YES	no	n/a	n/a	n/a	0	1	n/a	The Network's vision, values, and principles.
URL	foaf:homepage	non-literal	YES	no	n/a	n/a	n/a	0	1	n/a	The Network's homepage
Is made of	dc:terms:hasPart	non-literal	YES	no	no	no	YES	1	infinity	n/a	The SSE Initiative which the Network is made of
has address	essglobal:hasAddress	non-literal	YES	no	no	no	YES	0	1	n/a	The address of the network

FIG 3. Extract of DCAP-SSE Constrains Matrix



Studies in Metadata Practices— Session 1B

An Exploratory Analysis of Subject Metadata in the Digital Public Library of America

Hannah Tarver
University of North Texas
Libraries, USA
hannah.tarver@unt.edu

Mark Phillips
University of North Texas
Libraries, USA
mark.phillips@unt.edu

Oksana Zavalina
University of North
Texas, USA
oksana.zavalina@unt.edu

Priya Kizhakkethil
University of North Texas,
USA
priyakizhakkethil@my.unt.edu

Abstract

This paper presents results of an exploratory quantitative analysis of subject representation in the large dataset of over 8 million item-level metadata records in the Digital Public Library of America (DPLA) originating from a number of institutions that serve as content or service hubs of DPLA. The findings demonstrate both similarities and differences in subject representation across content and service hub providers. This benchmark study provides empirical data about the distribution of subjects at the hub level (e.g., minimum, maximum, and average number of subjects per record; number of records without subjects; and number of unique subjects) as well as distribution by hub type (content or service hubs), and subjects shared across similar hubs or across the entire aggregation.

Keywords: metadata aggregations; keywords; metadata values; subject analysis; subject terms

1. Introduction and Background

Cultural heritage institutions and funding agencies worldwide have invested intensively in digitization projects; however in many cases, access to those digitized collections often remained in separate pockets or silos. Large-scale digital libraries now bring together hundreds of individual digital collections and millions of items produced by these projects. The Digital Public Library of America (DPLA) is currently one of the most prominent such aggregations. Arising out of a vision from the early 1990s of a national digital library, shared by librarians, scholars, educators, and others, DPLA brings “different viewpoints, experience, and collections together in a single platform and portal, providing open and coherent access to our society’s digitized cultural heritage” (“About”, dp.la, 2015). Functioning on a distributed network model, DPLA consists of a group of national partners providing both content and services (Ma, 2014). DPLA was formed in 2010 and got underway in 2013 with support from a number of funding agencies which include the Alfred P. Sloan Foundation, the Arcadia Fund, the Institute of Museum and Library Services (IMLS), the John S. and James L. Knight Foundation, and the National Endowment for the Humanities (Mitchell, 2013).

Relying on a distributed network of partners to host and preserve digital information, DPLA focuses on the compilation of metadata to augment the discovery of these resources and to provide a useful platform where libraries and their patrons can make the best use of them. In addition, DPLA also provides APIs (Application Profile Interfaces) and maximally-open data to software developers, researchers, and others for building discovery tools along with providing access and communication (Ma, 2014). The DPLA community has also embraced the tenets of open data and adopted an advocacy stance in support of open access policies. On its launch in April 2013, a discovery platform provided access to an initial data set contributed by eighteen partners, or “hubs,” comprising more than two million records in over 3,200 collections. Since

the launch, the size of the aggregate collection and the number of partner institutions have continued to grow (Mitchell, 2013).

The internal data model of DPLA is based on the Resource Description Framework (RDF) and employs JSON-LD (JavaScript Object Notation-based serialization for Linked Data) for dissemination of metadata via API output. Based on the Europeana data model, the emphasis is on supporting the creation of graph structures and the standard is essentially a data aggregation and sharing service. Since the primary goal is the compilation of harvested data, some of the data gathered from providers is stored along with data generated or extracted during the data collection process. The DPLA metadata model is based on RDF and the central descriptive metadata standard employed is the Dublin Core (DC) (Mitchell, 2013). The metadata aggregated and normalized by DPLA is in the public domain and has no copyright restriction; DPLA data can be downloaded as JSON files, allowing for sharing or data analysis.

Although metadata analysis can lead in many directions, one field of significance is a subject field, since subject representation has applications in information retrieval, as well as in disciplines such as automated language processing and knowledge engineering that reference knowledge structures. In Svenonius (2000) definition, the “subject language” depicts what a document is about. Similarly, Soergel (2009) defines subject metadata in digital libraries as information concerning what the information object is about and why it is relevant.

Assigning subject metadata is based on subject analysis, for which various models have been proposed (e.g., Beghtol, 1986; Hjørland, 1998; Langridge, 1989; Šauperl, 2002; Wilson, 1968). These models guide the metadata creators to examine a document not only for its content, but also for author’s intentions, for viewpoints and possible bias, and to take into account when assigning subject terms the intended audience and intellectual level, as well as possible uses of information. According to Wilson (1968), since most works are multifaceted and cover more than one subject, the notion of “the” subject of a work is “indeterminate” (p. 318), i.e., in some cases it would be impossible in principle to decide between more than one different and equally precise descriptions to be the one and only subject of a work. Hjørland (1992) further developed this idea of multiplicity of a document’s subjects by taking the approach that subjects of a document can be defined as the informative or epistemological potentials of that document. According to Hjørland (1997), these intellectual potentials of a document can differ depending on periods of time and societal development, as well as across different domains, which would ideally require periodically revising subject headings in bibliographic records.

Subject metadata is crucial for providing access to information objects in both traditional library collections and digital collections and aggregations. To help achieve optimal recall and precision, it is recommended (e.g., ALCTS, 1999) to include Subject, Type, and Coverage elements in metadata records in digital libraries to accommodate different subject-related facets: topic, place, time period, language, etc. Gross & Taylor (2005) found that in the absence of subject headings in a catalog record, more than one third of the retrievals would be missed when a user performs a keyword search. In a study assessing the benefits of adding subject metadata to online records of the Northwestern University Library’s Eighteenth Century Collections Online (ECCO), Garrett (2007) extends the arguments forwarded by Gross & Taylor (2005) on the benefit afforded by subject headings for providing access even when the full text of a work is accessible. In a replication of the 2005 study, Gross, Taylor & Joudrey (2015) found that even with the addition of tables of contents and summaries or abstracts in the catalog records (which reduced lost hits), the absence of subject headings leads to an average of 27% of the retrievals to be missed.

Evaluation of metadata in digital libraries has gained more importance to ensure metadata quality (Hillmann, 2008). Margaritopoulos et al. (2009; 2012) discuss subject metadata from the point of view of measuring metadata quality, and in particular, completeness of metadata records. They point out that multivalued metadata fields such as subject are normally considered complete

if populated with at least one value; however multiple instances should be considered to determine the richness of the field, which can make the evaluation more complicated.

The empirical assessment of metadata has not yet become a common practice. In particular, few of the available studies that analyzed item-level metadata in digital libraries, included subject-metadata-related components. Several quantitative studies of item-level metadata in digital libraries (Jackson, Han, Groetsch, Mustafoff, and Cole, 2008; Kurtz, 2010; Weagley, Gelches, & Park, 2010) did not focus specifically on subject metadata but looked at the percentage of records that included one or more instances of each metadata element, including the subject metadata elements. For example, Kurtz's (2010) study of metadata in three university repositories revealed that the Dublin Core Subject field was included in only 65% of records. Weagley, Gelches, and Park's (2010) study of metadata in six digital video repositories reported the same level (65%) of Subject field utilization. To the contrary, Jackson and colleagues (2008) found Subject field values in almost all (94%) of metadata records harvested through OAI-PMH. The Dublin Core Coverage metadata element was found to be included in 7% and 21% of metadata records in the Kurtz (2010) Weagley, Gelches, and Park (2010) studies respectively and in 51% of records in the Jackson et al. (2008) study. Another study (Ma, Lu, Lin, & Galloway, 2009), which combined quantitative and qualitative approaches in overall analysis of item-level metadata in the Internet Public Library (IPL), evaluated users' ratings of the subject representation in IPL metadata through controlled-vocabulary subject headings and free-text keywords; the completeness of keywords was perceived to be quite low.

The analysis of literature reveals that little research to date has been conducted with the goal of specifically evaluating subject metadata in digital libraries. Available studies of subject metadata in digital libraries focused on collection-level metadata which describes entire collections of information objects as opposed to item-level metadata which describes each individual information object. For example, Zavalina (2011) examined and compared the free-text collection-level subject metadata (i.e., data values in the Description metadata field) across multiple digital libraries. The follow-up study (Zavalina, 2012) compared the data values in free-text Description and four controlled-vocabulary subject metadata fields -- Subjects, Temporal Coverage, Geographic Coverage, and Object Types/ Genres -- in three digital libraries: American Memory, Opening History, and The European Library. These two studies used a detailed manual content analysis and focused more on the qualitative characteristics of subject metadata than on quantitative ones. Some quantitative indicators that were measured in Zavalina (2012) study include the data value length (measured as the number of characters) -- range, median, mean, variance and standard deviation -- of each of the 5 subject metadata fields in the records.

The study reported in this paper is one of the first attempts to systematically evaluate subject metadata, and the first one to use a very large aggregator such as the Digital Public Library of America as its target.

2. Methods

The research questions that guided this exploratory study are: How are the subjects of information objects represented in metadata records across collections in the Digital Public Library of America (DPLA)? What are the differences and similarities in subject metadata originating from content hubs and service hubs?

Content hubs are digital repositories that maintain a one-to-one relationship with DPLA, providing metadata records for items owned or produced by that organization, such as ARTstor, California Digital University, The U.S. Government Publishing Office, and Harvard Library. Service hubs are state, regional, or other collaborative entities that bring together digital objects from multiple cultural heritage institutions and provide metadata records from all hosted or aggregated materials to DPLA through a single data feed. Some of the service hubs of DPLA are the Connecticut Digital Archive, Digital Library of Georgia, and The Portal to Texas History ("hubs", dp.la, 2015).

Unlike the previous studies of subject metadata in digital libraries that analyzed a generalizable sample of metadata records, the researchers of this study took a “big data” approach that analyzed the whole dataset and therefore avoided sampling errors. To address the research questions, the researchers used DPLA’s Bulk Download¹ to download the complete DPLA metadata dataset. This dataset was parsed into individual item records that contained both the original metadata from submitted by various DPLA hubs as well as a normalized version of the metadata in accordance with the DPLA Metadata Application Profile². In total the DPLA dataset (Phillips, 2015) contained 8,012,390 metadata records which were used in this analysis.

Each metadata record was parsed and the DPLA-normalized metadata was extracted for processing. The raw data for each field and the number of instances of the element in each record were added to a Solr index that the researchers used for their analysis in this paper; since the researchers chose to focus on subject terms for the purposes of this study, the data was limited to the dc:subject field values. Below is an example of the extracted and calculated data added to the Solr index for each field in the DPLA Metadata Application Profile for each record (Fig. 1). The example is represented in the JavaScript Object Notation (JSON) format that the researchers used for submitting data to the Solr index; this example shows that the record had two subject values, “Sun” and “Men.”

```
{
  "subject_ss": [
    "Sun",
    "Men"
  ],
  "subject_count_i": 2
}
```

FIG. 1. Example JSON created from a metadata record.

The researchers decided that for each record they would calculate the number of instances of each element in the record, and if there were no instances of that element in a given record then the count for that element would default to 0 for analysis.

The researchers used the Solr search framework to form queries for data analysis. Two components were particularly useful: the StatsComponent, which provides high level statistics for a specified field or set of fields in the index, and the Facet feature, which groups values, provides a count of instances of elements, and presents the number of records with a given value for a defined element. When the built-in features of Solr were not sufficient to answer the questions posed by the researchers, they wrote a series of Python scripts that would interact with Solr directly and apply additional logic and calculation to the data.

3. Findings

After general review of the data, the first finding of this analysis was that the average number of subjects per record in DPLA is 2.99, with a standard deviation of 3.90. In the dataset, 1,827,276 records had zero subjects, representing 22.8 percent of total records (see Table 1). For each hub, Table 1 lists the hub type, minimum and maximum number of subjects in the hub’s records, the number of items/metadata records, the total number of subject entries, the average number of subjects per record (mean), and standard deviation (stddev).

¹ <http://dp.la/info/developers/download/>.

² <http://dp.la/info/developers/map/>.

TABLE 1: Statistics for subject fields for each hub in the DPLA dataset.

Hub Name	Hub Type	Min	Max	Records	Subjects	Mean	Stddev
ARTstor	Content	0	71	56,342	194,948	3.46	3.47
Biodiversity Heritage Library	Content	0	118	138,288	454,624	3.29	3.41
David Rumsey	Content	0	4	48,132	22,976	0.48	0.69
Digital Commonwealth	Service	0	199	124,804	295,778	2.37	2.92
Digital Library of Georgia	Service	0	161	259,640	1,151,369	4.43	3.68
Harvard Library	Content	0	17	10,568	26,641	2.52	1.41
HathiTrust	Content	0	92	1,915,159	2,614,199	1.37	1.33
Internet Archive	Content	0	68	208,953	385,732	1.85	1.97
J. Paul Getty Trust	Content	0	36	92,681	32,999	0.36	1.21
Kentucky Digital Library	Service	0	13	127,755	26,009	0.20	0.78
Minnesota Digital Library	Service	1	78	40,533	202,484	5.00	2.66
Missouri Hub	Service	0	139	41,557	97,115	2.34	3.02
Mountain West Digital Library	Service	0	129	867,538	2,641,065	3.04	3.34
National Archives and Records Administration	Content	0	103	700,952	231,513	0.33	1.23
North Carolina Digital Heritage Center	Service	0	1,476	260,709	869,203	3.33	4.59
Smithsonian Institution	Content	0	548	897,196	5,763,459	6.42	4.65
South Carolina Digital Library	Service	0	40	76,001	231,270	3.04	2.35
The New York Public Library	Content	0	31	1,169,576	1,996,483	1.71	1.65
The Portal to Texas History	Service	0	1,035	477,639	5,257,702	11.01	4.97
United States Government Publishing Office	Content	0	30	148,715	457,097	3.07	1.75
University of Illinois at Urbana-Champaign	Content	0	22	18,103	67,955	3.75	2.87
University of Southern California Libraries	Content	0	119	301,325	863,535	2.87	2.67
University of Virginia Library	Content	0	15	30,188	95,328	3.16	2.33

This data showed some interesting results including that only the Minnesota Digital Library had at least one subject for all 40,533 of its records. There were two hubs, North Carolina Digital Heritage Center and The Portal to Texas History, which had individual records containing more than 1,000 subject headings (1,476 and 1,035 respectively). The average subjects-per-record ranged from 0.2 at the Kentucky Digital Library to 11.0 at The Portal to Texas History.

The next step was to break down the data based on hub types (service versus content hubs) for comparison (see Table 2). The researchers found that the average number of subjects for content hubs was 2.3 subjects per record, while the service hubs averaged 4.7 subjects per record. This means that service hubs tend to have twice as many subjects and keywords in their records as content hubs.

TABLE 2: Statistics for the subject field based on category (content hub or service hub).

Hub Type	Min	Max	Records	Subjects	Mean	Stddev
Content Hub	0	548	5,736,178	13,207,489	2.3	3.08
Service Hub	0	1,476	2,276,176	10,771,995	4.7	5.06

Further analysis of the metadata records originating from content hubs and service hubs showed that content hubs had a total of 1,590,456 records (28%) without any subjects compared to service hubs which had only 236,811 (10%) records without subjects.

The researchers also calculated additional metrics at the hub level for the DPLA records: the number of records without subjects, percentage of records without subjects, the mode of number of subjects-per-record, unique subjects, subjects unique to a single hub, and finally the entropy of the subject field for the specified hub (see Table 3). Entropy in this context represents a measure

of the average information content or similarity of values for a particular field, i.e., collections that have fewer unique values (more similar terms) will have a lower entropy score.

TABLE 3: Additional statistics for subject fields for each hub in the DPLA dataset.

Hub Name	Records	Records Without Subjects	% Without Subjects	Average Subjects per Record	Subject Count Mode	Unique Subjects	Subjects Unique to Hub	Entropy*
ARTstor	56,342	6,586	11.7	3.5	3	9,560	4,941	0.73
Biodiversity Heritage Library	138,288	10,326	7.5	3.3	2	22,004	9,136	0.65
David Rumsey	48,132	30,167	62.7	0.5	0	123	30	0.76
Digital Commonwealth	124,804	6,040	4.8	2.4	1	41,704	31,094	0.77
Digital Library of Georgia	259,640	3,216	1.2	4.4	2	132,160	114,689	0.67
Harvard Library	10,568	167	1.6	2.5	2	9,257	7,204	0.76
HathiTrust	1,915,159	525,874	27.5	1.4	1	685,733	570,292	0.88
Internet Archive	208,953	44,872	21.5	1.8	1	56,911	28,978	0.8
J. Paul Getty Trust	92,681	73,978	79.8	0.4	0	2,777	1,852	0.6
Kentucky Digital Library	127,755	117,790	92.2	0.2	0	1,972	1,337	0.62
Minnesota Digital Library	40,533	0	0	5	4	24,472	17,545	0.74
Missouri Hub	41,557	11,451	27.6	2.3	0	6,893	4,338	0.69
Mountain West Digital Library	867,538	49,473	5.7	3	1	227,755	192,501	0.68
National Archives and Records Administration	700,952	619,212	88.3	0.3	0	7,086	3,589	0.63
North Carolina Digital Heritage Center	260,709	41,323	15.9	3.3	2	99,258	84,203	0.66
Smithsonian Institution	897,196	29,452	3.3	6.4	7	348,302	325,878	0.62
South Carolina Digital Library	76,001	7,460	9.8	3	2	23,842	18,110	0.72
The New York Public Library	1,169,576	208,472	17.8	1.7	1	69,210	52,002	0.62
The Portal to Texas History	477,639	58	0	11	10	104,566	87,076	0.49
United States Government Publishing Office	148,715	1,794	1.2	3.1	2	174,067	105,389	0.92
University of Illinois at Urbana-Champaign	18,103	4,221	23.3	3.8	0	6,183	3,076	0.63
University of Southern California Libraries	301,325	35,106	11.7	2.9	2	65,958	51,822	0.59
University of Virginia Library	30,188	229	0.8	3.2	1	3,736	2,425	0.6

* Entropy calculated using the formula from Stivlia, Gasser, Twidale, Shreeves, & Cole (2004)

The data in Table 3 is helpful to identify hubs that have more coverage in the subject fields of their records. There is a range from the previously-mentioned Minnesota Digital Library that has zero records without subjects, or The Portal to Texas History that has 58 records (.01%) without subjects, to the National Archives and Records Administration with 88.3% and Kentucky Digital Library with 92.2% of their records lacking subject headings. The calculation of the number of subjects that are unique to a Hub showed that the Smithsonian Institution has 94% of its subjects unique to just the Smithsonian, while several other hubs share roughly half of their subjects with at least one other institution: ArtStor (52%), Biodiversity Heritage Library (42%), Internet Archive (51%), NARA (51%), University of Illinois at Urbana-Champaign (50%). The researchers theorize that the generally high number of unique subjects may be caused by the standard library practice of generating subject headings using the Library of Congress Subject Headings (LCSH); because of geographic and temporal qualification of the subjects, this creates a higher number of unique strings. Further analysis in this area could be performed to normalize LCSH into its constituent pieces and re-run the analysis to determine what effect this has on the dataset.

The researchers compiled the same information by hub type (see Table 4) to analyze the overlap of subject terms between hubs of different types.

TABLE 4: Makeup of unique subjects per hub type in the DPLA.

Hub Type	Records	Unique Subjects	Subjects Unique to Hub Type	% of Subjects Unique to Hub Type
Content Hub	5,736,178	1,311,830	1,253,769	96
Service Hub	2,276,176	618,081	560,049	91

A large percentage of subjects -- 96% for content hubs and 91% for service hubs -- are unique to that hub type. In fact, only 3% of the total unique subjects in the dataset are shared between content hubs and service hubs (see Fig. 2).

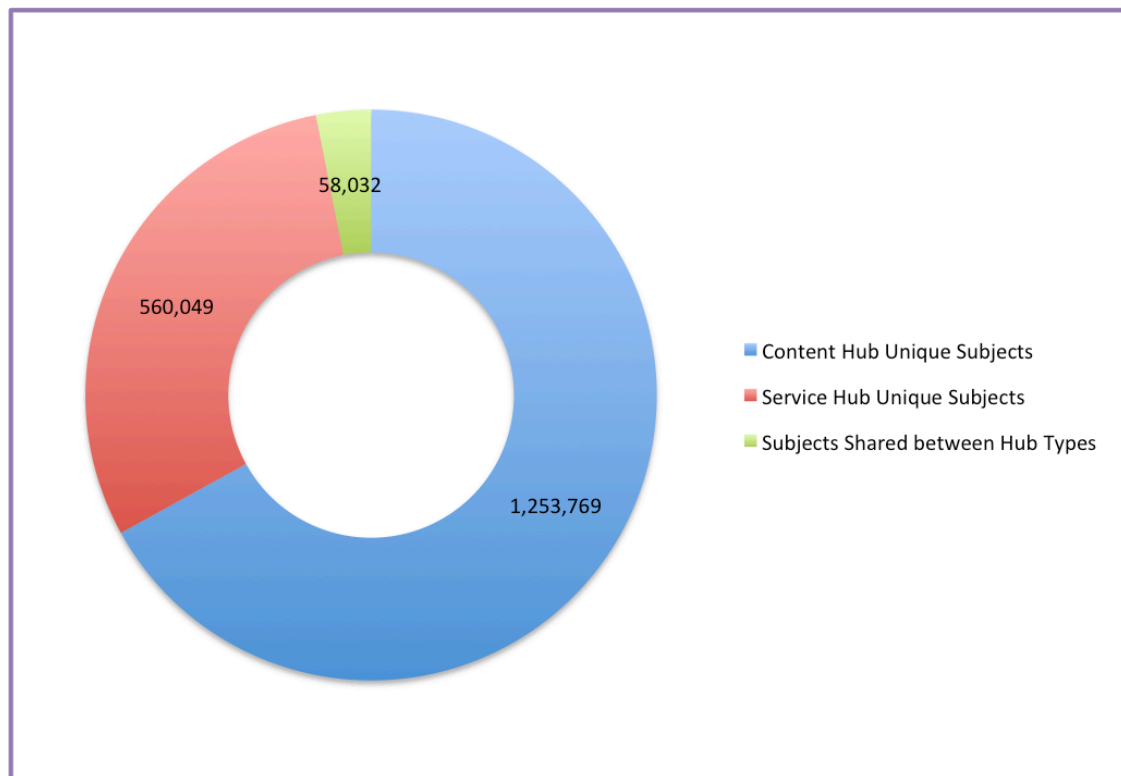


FIG. 2. Comparison of unique and shared subjects between hub types in DPLA.

The next step was to look at shared subjects, which is significant since subject terms have a relatively unique ability to connect users with disparate resource types, and across multiple partner collections, that have common topical content. However, this assumes a level of consistency in subject assignment, so the analysis determined how many subjects were shared across individual hubs and the subjects common to the highest number of hubs (see Table 5).

TABLE 5: Subjects shared, by number of hubs.

Unique Subject Count	# of Hubs with Subject	Unique Subject Count	# of Hubs with Subject
1,717,478	1	302	12
114,047	2	245	13
21,126	3	199	14
8,013	4	152	15
3,905	5	117	16
2,187	6	63	17
1,330	7	62	18
970	8	32	19
689	9	20	20
494	10	7	21
405	11	7	22

Table 5 demonstrates that a large majority of subjects (roughly 92%) are unique to a single hub. Subjects that are shared between two hubs represent 6% of total subjects and only 1% are shared among three hubs. The total number of remaining subjects, shared among four or more hubs, amounts to only 1.5% of total subjects.

The seven subject headings that are shared by twenty-two hubs are: African Americans, Animals, Architecture, Children, Education, Horses, and Transportation.

4. Discussion and Conclusions

Subject terms have a unique place in metadata for several reasons. First, every item has one or more “topics,” or content that can be described in topical ways, so it is reasonable to expect that complete metadata records should include subject terms, or that records without terms could be updated given time and resources. This differs from many other fields in a metadata record, for which entries may remain blank simply because the information (e.g., creator, location, etc.) is not known about the item. Secondly, although many metadata fields may be complete with a single data value (e.g., creation date or item language), subject fields often occur as multiple entries in each record, and in most cases, additional subject terms are directly related to additional access points for users (i.e., providing more subject terms increases the findability and usefulness of a metadata record). Finally, to some degree, subject representation requires a certain level of active consideration – that is, a metadata creator has looked at the item, thought about the content, and then generated or assigned subject terms. This suggests that data values associated with subject fields in metadata records can often be tied to curation activities within individual hubs, as opposed to data values in other fields of metadata records which *may* be copied directly from the source item or from accompanying collection-level information (e.g., book titles, or creator names).

This analysis provides a framework for general discussion regarding subjects in digital collections, and in large aggregates. One noticeable finding is the high variability of the number of instances of subject fields across records, ranging from no subjects to more than one thousand. Reasons for these variances would have to be explored locally at individual hubs – for example, records that do not have any subject terms may be due to workflow issues, a lack of tools to discover incomplete records or resources to fix known deficits, or even local practices that do not require or encourage subject representation. Several hubs also had records containing a large number of subject terms (i.e., more than 100, more than 500, or more than 1,000). Based on the experiences of the researchers handling records in The Portal to Texas History, some of the numbers may be slightly inflated due to the normalization process that DPLA uses when importing records. For example, the Portal has a locally-established hierarchical subject vocabulary, the UNT Libraries Browse Subjects (UNTL-BS), that is parsed into separate

keywords when records are harvested and added to DPLA; for example the hierarchical subject string “Business, Economics and Finance - Transportation - Automobiles” becomes keywords “Business, Economics and Finance,” “Transportation,” and “Automobiles.” This means that a record with only one or two controlled terms from the UNTL-BS list may have six or eight keyword terms in the DPLA-normalized record. While this may not account for the extremely large numbers, it could impact some hubs more than others. Additionally, most of the records in the Portal containing higher numbers of subject terms tend to contain many personal names. In fact, the outlier in this dataset is a metadata record representing a ledger of inquest records for which the partner particularly requested that all of the names be included in the metadata (since the ledger is handwritten).

Another discussion point arising from this analysis is the differences in average number of subject terms between hub types. DPLA content hubs provide more than double the number of records that service hubs contribute, however the average number of subjects per record for content hubs is half that of the service hubs. This may be related to the fact that service hubs aggregate or host materials from multiple institutions, and therefore the initial metadata creation or maintenance may be distributed among content holders. Overall, the numbers show a large amount of variance even among hubs of the same type, so it is hard to say with certainty if the differences are more representative of an actual divide by hub type, or of radical differences among individual hubs.

While determining the accuracy and “quality” of subject metadata in these records would be essentially impossible on a large scale, this analysis does provide data related to completeness, i.e., whether or not all records have subject(s), assuming that every record should include at least one subject term. It also highlights those metadata records that do not fit the model of an average record within a particular digital library and may be indicative of problem records or lower quality metadata. On a local level, subject analysis similar to the analysis presented in this paper could help individual hubs to discover gaps or possible areas of metadata enhancement within their own collections. Some examples include identifying records that have no subject entries or for which the number of unique values is higher or lower than expected for the known content.

Aside from records in individual hubs, the findings also highlight the lack of overlap across all of the collections in DPLA since the majority (92%) of subject terms in metadata records are unique to a particular hub. While some of this uniqueness in subject terms might be explained by uniqueness of items contributed to DPLA by individual hubs and varying subject matter of these items, this factor would only contribute a single-digit percentage of uniqueness of subject terms in DPLA. It is likely that most of the 92% uniqueness is due merely to the lack of a common controlled vocabulary. Since DPLA aims to bring items together for access, using fewer unique subject terms across DPLA would appear to be of importance to facilitate finding and collocating materials across the aggregate. However, implementing any plan to improve consistency in subject representation across such a large number of records and content providers would be difficult, time consuming, and could require extensive resources as well as buy-in from the many hubs. Perhaps one option based on the kind of analysis in this paper would be to provide better access to currently-used or most-used subject terms in DPLA metadata for persons who maintain records at individual hubs. While it would not be an immediate fix, it could create an opportunity to start promoting intentional subject overlap.

4.1. Further Study

As this study has shown, the availability of data from DPLA creates an opportunity for various kinds of metadata analysis across aggregated collections or at local institutions. Additional analyses of subject representation in DPLA could look at field values across the collection after basic normalizations. For example, known Library of Congress Subject Heading (LCSH) terms could be broken into constituent pieces in the same way that OCLC parses values into Faceted Application of Subject Terminology (FAST) terms (e.g., “Children--Texas” into “Children” and “Texas”). This could show whether a larger overlap in subject matter exists than is apparent from

analysis of original subject strings. Qualitative studies could also provide context regarding the data in this study, such as the reasons that some records have no subject terms, the differences in the number of subject terms across hubs or hub types, and additional information about the lack of overlap in subject terms within DPLA.

In addition to the dc:subject metadata field, several other fields particularly lend themselves to cross-collection analysis at an aggregate level. For example, coverage field(s) function similarly to subject in the way that they represent content of materials. Analysis of dates, time periods, and geographic elements in coverage values could show where topics converge, or where information could be easily added to provide more item-level access.

On an even broader scale, comparisons of DPLA with other large international digital libraries or aggregates (such as Europeana, Canadiana, etc.) would provide a more extensive dataset for studies in metadata completeness or metadata field usage. The data in this study provides a baseline that could be used as a point of comparison regarding subject term representation in individual metadata records or overlap across large collections and aggregates.

References

- ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. (1999). Subject Data in the Metadata Record: A Report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis Working Draft. Retrieved from <http://archive.ifla.org/VII/s12/mom/appendx3.htm>
- Beghtol, Claire. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42(2), 84-113.
- Digital Public Library of America. (n.d.). Retrieved March 20, 2015, from <http://dp.la/>.
- FAST (Faceted Application of Subject Terminology). (2013, August). Retrieved April 1, 2015 from <http://www.oclc.org/research/themes/data-science/fast.html>.
- Garrett, Jeffrey. (2007). Subject headings in full-text environments: The ECCO experiment. *College & Research Libraries*, 68(1), 69-81.
- Gross, Tina, and Arlene G. Taylor. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, 66(3), 212-230.
- Gross, Tina, Arlene G. Taylor, and Daniel N. Joudrey. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly*, 53(1), 1-39.
- Hillmann, Diane Ileana. (2008). Metadata quality: from evaluation to augmentation. *Cataloging & Classification Quarterly*, 46(1), 65-80.
- Hjørland, Birger. (1992). The concept of 'subject' in information science. *Journal of Documentation*, 48(2), 172-200. doi: 10.1108/eb026895
- Hjørland, Birger. (1997). The concept of subject or subject matter and basic epistemological positions. In *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport CT: Greenwood Press, 55-103.
- Hjørland, Birger. (1998) Theory and metatheory of information science: a new interpretation. *Journal of Documentation* 54, 606-621.
- Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustaffoff, and Timothy W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata*, 8(1), 5-21.
- Kurtz, Mary. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology & Libraries*, 29(1), 40-46.
- Langridge, Derek Wilton. (1989). *Subject analysis: principles and procedures*. London: Bowker-Saur.
- Ma, Hong. (2014). Techservices on the Web: DPLA: Digital Public Library of America. *Technical Services Quarterly*, 31(1), 83-84. doi: 10.1080/07317131.2014.845013
- Ma, Shanshan, Caimei Lu, Xia Lin, and Mike Galloway. (2009). Evaluating the metadata quality of the IPL. *Proceedings of the Annual Meeting of American Society for Information Science and Technology*, 49. <http://www.asis.org/Conferences/AM09/open-proceedings/papers/49.xml>
- Margaritopoulos, Thomas, Merkourios Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. (2009). A fine-grained metric system for the completeness of metadata. In Fabio Sartori, Miguel-Angel Sicilia, & Nikos Manouselis (Eds.), *Metadata and semantic research* (pp. 83-94). Berlin: Springer.

- Margaritopoulos, Merkourios, Thomas Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. (2012). Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology*, 36(4), 724–737. doi:10.1002/asi.21706.
- Mitchell, Erik T. (2013). Three case studies in linked open data. *Library Technology Reports*, 49(5), 26-43.
- Phillips, Mark Edward. (February 2015). Digital Public Library of America: Bulk Metadata Download, February 2015 [dataset]. <http://digital.library.unt.edu/ark:/67531/metadc502991>
- Šauperl, Alenka. (2002). Subject determination during the cataloging process: observation. Lanham, MD: Scarecrow Press.
- Soergel, Dagobert. (2009). Digital libraries and knowledge organization. In S. R. Kruk & B. McDaniel (Eds.), *Semantic Digital Libraries*, (pp. 9-39). Berlin: Springer.
- Svenonius, Elaine. (2000). *The intellectual foundations of information organization*. Cambridge, MA: MIT Press.
- Stvilia, Besiki, Les Gasser, Michael B. Twidale, Sarah L. Shreeves, and Tim W. Cole. (2004). Metadata quality for federated collections. In S. Chengulur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the International Conference on Information Quality — ICIQ 2004* (pp. 111–125). Cambridge, MA: MIT.
- Weagley, Julie, Ellen Gelches, and Jung-Ran Park. (2010). Interoperability and metadata quality in digital video repositories: A study of Dublin Core. *Journal of Library Metadata*, 10(1), 37-57.
- Wilson, Patrick. (1968). *Two kinds of power: An essay on bibliographic control*. Berkeley: University of California Press.
- Zavalina, Oksana L. (2011). Free-text collection-level subject metadata in large-scale digital libraries: A comparative content analysis. In T. Baker, D. I. Hillmann & A. Isaac (Eds.), *Proceedings of the International Conference on Dublin Core and Metadata Applications*, (pp. 147-157). The Hague: Dublin Core Metadata Initiative.
- Zavalina, Oksana L. (2012). Exploring the richness of collection-level subject metadata in three large-scale digital libraries. *International Journal of Metadata, Semantics, and Ontologies*, 7(3), 209-221. doi: 10.1504/IJMSO.2012.050182

Exposing Library Holdings Metadata in RDF Using Schema.org Semantics

Myung-Ja K. Han
University of Illinois at Urbana-
Champaign, USA
mhan3@illinois.edu

Timothy W. Cole
University of Illinois at Urbana-
Champaign, USA
t-cole3@illinois.edu

Patricia Lampron
University of Illinois at Urbana-
Champaign, USA
lampron2@illinois.edu

M. Janina Sarol
University of Illinois at Urbana-
Champaign, USA
mjsarol@illinois.edu

Abstract

Libraries have been busy transforming and publishing their data as linked open data by testing already existing semantics and developing new sets of semantics. So far, most of the efforts have focused on the bibliographic data, not the holdings and item related data that are unique to individual libraries and that help users access the information resources they need. The University of Illinois at Urbana-Champaign Library experimented with a subset of its bibliographic records (5.4 million) describing print resources and associated holdings data to examine options and best practices so far identified for expressing library holdings data using schema.org semantics. The experimentation suggests that the mappings for holdings data recommended by the BibExtend Community Group are in some ways incomplete and that some proposed uses of schema.org types and properties to describe library holdings go beyond current schema.org definitions. Existing schema.org enumerations should be extended (e.g., regarding availability) to better describe library use cases, and some extensions to schema.org are needed to fully describe library holdings data and to maximize their utility. This paper highlights issues, suggests potential extensions identified during the transformation to schema.org semantics, and discusses options to make essential library holdings data fully visible as linked open data.

Keywords: Linked Open Data; library catalog; holding data; schema.org; Semantic Web

1. Introduction

Libraries today are both producers and consumers of linked open data (LOD). In describing library resources, libraries need to identify what unique information they can and want to contribute to the growing Web of Data (aka the Semantic Web) and to assess which semantics will be most effective for sharing resource descriptions. Their role as consumers of LOD can help inform these decisions. To date, libraries have tried and tested a variety of data models and semantics to publish catalog records as linked data (Cole, Han, Weathers, & Joyner, 2013). Two initiatives, the Library of Congress (LC)' BIBFRAME (Library of Congress, 2015) and schema.org (schema.org, 2015) as used, for example, by the Online Computer Library Center (OCLC) (OCLC, 2014) have garnered the majority of interest. The graphs produced by transforming library catalog records to BIBFRAME or schema.org are useful, but incomplete; less attention has been given so far to holdings data, which is essential to help users know where to locate information resources and how to access them. This is because libraries maintain holdings data separate from their bibliographic descriptions, e.g., in acquisitions and circulation modules in Integrated Library System (ILS) or Electronic Resource Management (ERM) systems.

Using a snapshot of the University of Illinois at Urbana-Champaign (UIUC) Library's print bibliographic and holdings data, this paper examines options and best practices so far identified for expressing library holdings data using schema.org semantics. Web search engine vendors

collaborated to create schema.org, and the perspective is decidedly commercial. Preliminary mappings of holdings data to schema.org have been proposed, notably by the W3C BibExtend Community Group (Schema BibExtend Community Group, 2015), but our examination suggests that these mappings are incomplete and some proposed uses of schema.org types and properties (i.e., Resource Description Framework (RDF) classes and predicates) go beyond current definitions. Enumerations are insufficient in a few cases (e.g., regarding availability and borrowing terms), and extensions to schema.org are needed to fully describe library holdings data. We also found that the holdings data contained in our ILS acquisitions and circulation modules, while adequate to generate RDF descriptions of print holdings, were not adequate to generate RDF descriptions of electronic holdings. In this paper, we highlight issues identified from the experimentation, suggest potential extensions to schema.org semantics, and discuss options libraries may want to pursue to make their holdings data visible as LOD. This work remains incomplete and further research is ongoing. For example, conflation of work, expression, manifestation, and item data makes matching and de-duping across collections difficult, but OCLC's Work Identifiers (OCLC, 2015) may provide a solution.

2. Library Holdings Data

2.1. Holdings Data in MARC

Libraries have been using the MACHine Readable Cataloging (MARC) format as a cataloging tool since the early 1960s. Although both bibliographic and holdings data can be encoded in MARC (Library of Congress, 2015), most ILS manages bibliographic and holdings data in different modules. Because of this, the term 'library MARC record,' usually refers only to the bibliographic data without the holdings data.

A critical history of English literature.
Main Author: Daiches, David
Published: New York, Ronald Press Co. [1960]
Topics: English literature - History and criticism
Tags: No Tags, Be the first to tag this record! Add record!

More Details | **Location & Availability** | User Reviews | Request Item

University of Illinois at Urbana-Champaign

Location: Literatures & Languages
Call Number: PR93 .D29 1960
Text me this call number

Copy: 2
Library Has (Volumes): v.1 c.2
v.2 c.2
Status: Available

Location: Main Stacks
Call Number: 820.9 D14C
Text me this call number

Copy: 3
Library Has (Volumes): v.1 c.3
v.2 c.3
Status: Available

Location: Oak Street Facility [request only]
Call Number: 820.9 D14C
Text me this call number

Copy: 4
Library Has (Volumes): v.1 c.4
v.2 c.4
Status: Available

FIG. 1-1.

Meditationes emblematicae de restaurata pace Germaniae = Sinnbilder von dem widergebrachten Teutschen Frieden /
kürzlich erklärt durch Johann Vogel.

Main Author: Vogel, Johann
Other Names: Zunner, Johann David,
Published: Francofurti : Apud Joh. Dav. Zunnerum, [1649]
Topics: Emblems - Early works to 1800. | Emblem books, Latin - Germany - 17th century. | Emblem books, German - Germany - 17th century.
Genres: Emblem books - Germany - 17th century.
Online Access: Full text - UIUC
Online Access: Full text - OCA
Tags: No Tags, Be the first to tag this record! Add

More Details | **Location & Availability** | User Reviews | Request Item

University of Illinois at Urbana-Champaign

Location: *UIUC Online Collection
Call Number: Online Resource
Text me this call number

Copy: 1
Online Access: Full text - UIUC
Online Access: Full text - OCA

Location: Rare Book & Manuscript Library [non-circulating]
Call Number: Emblems 0075
Text me this call number

Copy: 1
Related Information: Request the item for use in the Rare Book & Manuscript Library
Status: Available

FIG. 1-2.

FIG. 1. Holdings and item specific data displayed in the UIUC Library's OPAC. Figure 1-1 shows the multipart items and Figure 1-2 shows the print and online holdings shown together.

"Copy-specific information for an item; information that is peculiar to the holding organization; information that is needed for local processing, maintenance, or preservation of the item; and version information" are collectively referred to as holdings data. Holdings are sub-classed into three different types "single-part, multipart, or serial item (Library of Congress, 2006)," and each copy will have one holdings record, i.e., there are three serial holdings if there are three copies of

the title. Not encoded in MARC, the ILS also has circulation or status of the item data in the system. In addition to these three traditional types of holdings, we also consider items with both print and electronic holdings as a new class. Figure 1 illustrates the types of holdings (and item status) data as displayed in the UIUC Library's Online Public Access Catalog. As shown in Figure 1-1 for the title holdings information of David Daiches' *A Critical History of English Literature*, users can see information about the item's location, the call number, copy number, and available volume information specific to the copy of the title (as well as availability and the status of the item which are not normally encoded in MARC). As illustrated in Figure 1-2, a link to an online copy of the item is also provided if it is available. A barcode of each item is also available in the ILS system but is not displayed to users, because it is not used for the search. According to our analysis of types of holdings data, while a majority (72%) of the titles in the sample set of 5.4 million records are associated only with a single copy of a single-part holding, multiple copies and other holding classes are also represented, and 6% of the titles have both print and online holdings as shown in Figure 2. (Note, we treated links to the full contents included in the data field 856 (Electronic Location and Access) as additional online holdings.)

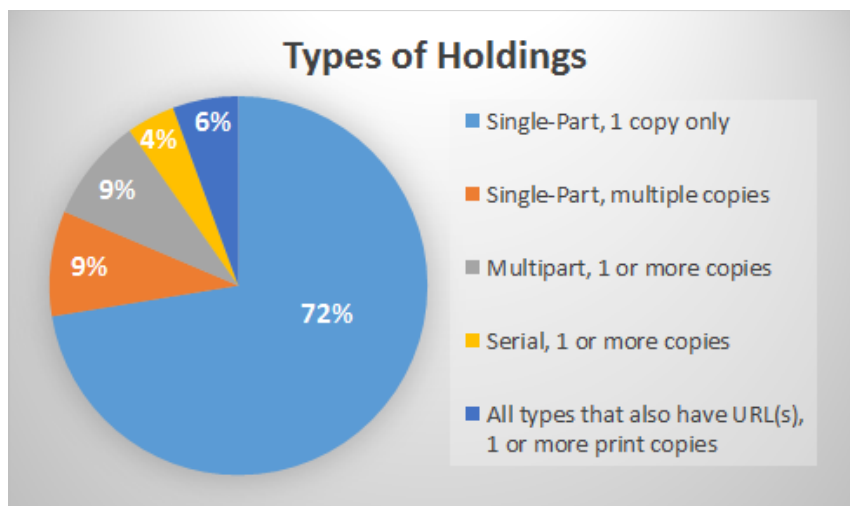


FIG. 2. Types of Holdings associated with UIUC Print Catalog Records

2.2. Relationship Between Bibliographic and Holdings Records

While bibliographic data describes manifestation and higher level information according to the Functional Requirements for Bibliographic Records (FRBR)'s group 1 entities (Tillet, 2004), the holdings data includes both manifestation and item level information. Because one manifestation can link to more than one item, the relationships between bibliographic data, holdings data, and item data may vary depending on the number of copies each library has, or the type of resource, e.g., whether the resource is a monograph, multipart item or serial. For example, one bibliographic record can have more than one holdings record, and each holdings record can have more than one related item with item specific information, such as a barcode and coverage (or volume) information if the item is part of a serial or multipart item. So the relationship between bibliographic record and holdings record(s) is one-to-one or one-to-many, and the relationship between holdings record and item data is also one-to-one or one-to-many. (If an item is a part of series, then the item data may have two different bibliographic records, one that describes a series and the other that describes the specific item.)

Traditionally, the library manages holdings data at the copy level, i.e., the unit of the holdings data is based on the copy number. For example, if the copy is for a multipart item or serial, volume specific information is organized and added under the copy level. So if a user wants to know the availability of a specific issue or volume of the serial or multipart item, the user has to

check the item at copy level first, and then check the volume information from the next layer of the information structure.

3. Holdings Data in schema.org

OCLC decided to use schema.org semantics for expressing bibliographic metadata in RDF for the simple reason that the majority of search engines support schema.org, which means these resources are more easily searchable and discoverable on the web. More than 97% of UIUC's sample set of 5.4 million MARC bibliographic records include OCLC numbers. This allowed us to focus on how best to describe holdings data, i.e., data unique to UIUC that can be easily integrated with the bibliographic LOD graphs already published by OCLC. We acknowledge that the Document Availability Information (Voß and Reh, 2015) group has been working on describing holdings as LOD and developed its own ontology. However, we decided to use schema.org in line with bibliographic data already available in OCLC in order to improve discoverability of holdings data on the web. To create schema.org RDF, we used two transformation stylesheets; a modified version of LC's transformation stylesheet (Library of Congress, 2014) to transform MARCXML records with holdings data to Metadata Object Description Schema (MODS), and a locally created transformation stylesheet to transform MODS to schema.org RDF. We decided to use MODS as a transit metadata format because the MODS top element <location> can properly contain library holdings data, and the schema allows us to include linked data source URIs as values.

Our starting point for this work was the recommendation offered by the BibExtend Community Group (2014). For cases where a bibliographic record is associated with a single copy of a single-part print item (holding), the BibExtend Community Group recommendation works well (see Table 1). With modest modification the recommended "Holdings via Offer" approach allowed us to express most of the critical information contained in this type of holdings data. For instance, we can express the holding organization (schema predicate seller) and branch (availableAtOrFrom) using the Offer class. The IndividualProduct class allows us to express the barcode (serialNumber) and call number (sku). We do, however, deviate from the BibExtend Community Group's recommendation in some particulars; these are further explained in Section 5 below.

TABLE 1: Similarities and differences between the BibExtend Community Group's recommendations and UIUC approaches in mapping holdings and item specific information to schema.org semantics

Holdings and item specific information	Schema.org/BibExtend Community Group Recommendation	UIUC Approaches
Library	Seller	seller
Shelving Location	availableAtOrFrom	availableAtOrFrom/Place/name
Call number	Sku	itemOffered/IndividualProduct/sku
Item barcode	serialNumber	itemOffered/IndividualProduct/serialNumber
Copy number	Not mapped	itemOffered/IndividualProduct/name
Borrowing terms	businessFunction	Not mapped
Item status	availability	Not mapped
Online Holdings	Not mapped	itemOffered/IndividualProduct /url

4. More Complex Holdings Scenarios

The purpose of exposing the library's holdings data as LOD is to allow users to find, identify, select, and obtain (IFLA, 1998) the exact information resource they need even when it is only available in print. Holdings information can also be a way to filter search results. As illustrated in Figure 2, bibliographic data can be linked to single-part, serial, or multipart holdings as well as to online holdings. A library may hold multiple copies in one or more locations. An item may be available both in print and online as a scanned set of page images. A monographic resource may

have been published in multipart volumes (e.g., a 19th-century triple-decker novel). Serial bibliographic entities are published over time, one issue or volume at a time. Users need access to both the bibliographic information and to exactly which issues and volumes are available from an institution. It is not enough to know that ten libraries in the country own at least some volumes of an obscure journal; a user wants to know which of the ten libraries own volume 4 (the volume that the user needs).

```

schema:hasPart [
  a schema:PublicationVolume ;
  schema:volumeNumber "1972" ;
  schema:offers [
    a schema:AggregateOffer ;
    schema:seller <http://id.loc.gov/authorities/names/n79066210> ;
    schema:itemOffered [
      a schema:SomeProducts ;
      schema:offers [
        a schema:Offer ;
        schema:availableAtOrFrom [
          a schema:Place ;
          schema:name "Oak Street Facility [request only]"
        ] ;
        schema:itemOffered [
          a schema:IndividualProduct ;
          schema:sku "324.23 Un3m" ;
          schema:serialNumber "30112071980053" ;
          schema:name "Copy Number: 1"
        ]
      ]
    ], [
      a schema:Offer ;
      schema:availableAtOrFrom [
        a schema:Place ;
        schema:name "Oak Street Facility [request only]"
      ] ;
      schema:itemOffered [
        a schema:IndividualProduct ;
        schema:sku "324.23 Un3m" ;
        schema:serialNumber "30112063348632" ;
        schema:name "Copy Number: 2"
      ]
    ]
  ] ;
  schema:offerCount "2"
] ;
schema:offers [
  a schema:Offer ;
  schema:seller <http://id.loc.gov/authorities/names/no2015002503> ;
  schema:itemOffered [
    a schema:IndividualProduct ;
    schema:url <http://purl.access.gpo.gov/GPO/LPS6982> ;
    schema:description "electronic resource"
  ]
]

```

FIG. 3. Multipart library print and electronic holdings information serialized with schema.org semantics.

To accommodate complex holding scenarios involving multiple copies (i.e., multiple holdings), we employed schema.org's AggregateOffer and SomeProducts types. These were suggested in the BibExtend Community Group recommendation as possibly being of use when describing consortial holdings, but we found them useful for our single institution to deal with multiple copies in different department libraries. This approach also anticipates aggregation of

holdings LOD from multiple sources. Multiple holdings from the same institution— print and local digital copies – are grouped within the same `AggregateOffer`. Using `AggregateOffer` allows us to provide the number of copies from each institution through the `offerCount` property. Local online holdings are also included within an `Offer` rather than using the `url` property under the `CreativeWork` class in order for descriptive information about the electronic copy to be included. For monographic resources published in multiple volumes (multipart) and serial publications, we used the `hasPart` property under the `CreativeWork` class. Each volume has a type of `PublicationVolume`, which allows the enumeration and chronology to be specified using the `volumeNumber` property. Multiple copies of a single volume appear within an `AggregateOffer` description. This usage of `AggregateOffer` and `SomeProducts` is illustrated in Figure 3.

5. Discussion and Recommendations

5.1. Variations from BibExtend Community Group Recommendations

Our explicit use of `IndividualProduct` deviates from the BibExtend Community Group recommendation to use `sku` and `serialNumber` predicates under `Offer` class, which is a shortcut way to express the serial number and barcode of the product implied in the `Offer`. Having `Offer` as the domain for these properties created problems when we looked at more complex holdings examples. Our bibliographic records with multiple holdings consist of several products, so we decided not to add `Product` as an `additionalType` property to the bibliographic record. Instead, we decided to define each item or digital instance as an `IndividualProduct`.

Another way that we deviated from the BibExtend Community Group's approach was by not including the borrowing terms (`businessFunction`) or the item status (`availability`). In both cases we felt the enumeration of possible values for these predicates was insufficient (see below). In addition, we did not include the borrowing terms, which the BibExtend Community Group recommends changing to `LeaseOut` (<http://purl.org/goodrelations/v1#LeaseOut>), because some of our holdings are not eligible to be loaned (e.g. non-circulating items) and the definition of `LeaseOut` does not adequately describe library loan policies.

Because it provides additional identifying information about print volumes, we include the copy number in the `IndividualProduct` description, something not anticipated by the BibExtend Community Group recommendations.

5.2. Challenges of Working with Holdings Data

Gathering Holdings Data from Various Sources: Information about availability (schema property) is difficult or impossible to acquire through static holdings data, and requires cooperation with a more dynamic and live data source, such as a circulation database. For online access to multipart items and serials, it has proven difficult to express which resources are available through which services and the coverage of the serials. For instance, the `Offer` information for a journal with electronic issues divided by provider would require harvesting that information outside of the traditional bibliographic and holdings data in an ILS, possibly through close work with vendors or through a separate ERM system maintained elsewhere in the library, in order to correctly display this availability information to users.

Irregular Formatting of Volume Information: Some multipart/serial enumeration and chronology fields in holdings data are irregularly formatted. The value may be a volume, year, or another pattern. For example, one holding may contain volumes 1 and 2 of a serial publication, while another holding for the same record only contains volume 1. In some cases, the information can be completely different from others based on historic binding decisions. This makes it difficult to share serial holdings information across institutions, sometimes even across branches within the same institution. However, we think that this kind of string-based practice can be corrected easily by assigning a permanent identifier for each volume when the item is published.

Data Created Using Different Practices: The way libraries create catalog records has changed over the years, and because of this, catalogs often contain various bibliographic records that follow different rules and standards. For example, the UIUC catalog records include bibliographic records that describe manifestation and expression. The UIUC library has a single record approach when a print book's digital copy is available in open access and the Library has not locally digitized it, i.e., the print record has a url of the digital copy. However, the Library creates a separate record for all purchased electronic and in-house digitized books (separate record approach). The separate record approach can result in disconnected CreativeWork descriptions (one linked to a print item Offer, and one linked to an online Offer) for essentially the same intellectual content. On the other hand, the single record approach results in some CreativeWork descriptions linking simultaneously to print item and online Offers, although the bibliographic data only describes the print copy.

5.3. Representing Holdings Data into RDF

The library manages holdings data at the copy-level and provides available volumes in the next layer. For our experimentation, when we transformed our records into RDF, we changed this relationship. We mapped the holdings data at the volume-level (using hasPart for serial and multipart items), and provided the copy information in the next layer. We think this approach benefits users by allowing them to find the item related data directly from the bibliographic data without searching from the copy level data. The volumes available to a user can be easily expressed with this approach. However, we recognize that this doesn't mean that libraries have to expose their entire holdings and item related data as LOD on the web since not all data that libraries use to manage and organize their resources are beneficial for discovery and access. In addition, some holdings data is not easy to integrate with the data in the ILS and in MARC format.

5.4. Limitations of schema.org For Expressing Holdings

Because schema.org is designed to accommodate structured commercial data, there are instances where schema.org semantics do not align conveniently with library data.

Immediate Availability: While the BibExtend Community Group recommends expressing item availability using: InStock, OutOfStock, PreOrder, or InStoreOnly, this does not capture all possible information about the item status and availability used in the library. It would also be beneficial to our users to further provide availability and accessibility data by adding information describing loan periods or class reserves, but providing this data requires new properties. Additionally terms like OutOfStock do not really describe that an item is currently loaned out and expected back at the end of the loan period. One practitioner has suggested using availabilityStarts to indicate when an item is expected back from loan (Scott, 2014). We recommend better enumeration values for item availability, for example, AvailableToLoan, OnLoan, and InLibraryUseOnly (or RoomUseOnly). We also recommend adding more information such as how long an item can be loaned (eligibleDuration), and for electronic holdings, until when an item can be accessed (validThrough).

Borrowing Terms: The BibExtend Community Group recommends adding the businessFunction property with a value of LeaseOut (<http://purl.org/goodrelations/v1#LeaseOut>) to describe that a library item is available to be borrowed. While better than the default value, which is for sale, this does not adequately describe library loans, nor does it account for items that cannot be loaned (e.g., in-library use only). We recommend adding more enumerations to borrowing terms (Loan, NonCirculating, Request).

Eligible Customers: Print loan requires a current and valid library ID, and may also require additional conditions be met, e.g., enrollment in class. Customer type may also dictate the loan period or other access constraints. Further complicating the issue of online access is the conditions of use prescribed by various vendors, e.g., requiring a campus IP address, VPN connection, or number of concurrent users. The eligibleCustomerType property in schema.org

expects it to be of type `BusinessEntityType`, which can have values of `Business`, `Enduser`, `PublicInstitution`, or `Reseller` from the `GoodRelations` (<http://purl.org/goodrelations/v1>) data model. These enumerations are inadequate to describe such information, and the other requirements such as having a library ID or a login account cannot be described. We propose adding an `Offer` property (requires) to describe the eligibility requirements.

6. Conclusions

Complete holdings data is essential for the user to locate and obtain/access the precise representation or component of the item that the bibliographic data describes, and this unique holdings and item related data can be provided only by an institution that holds the particular information resources. The UIUC Library's research on exposing library holdings data as LOD revealed that holdings data has unique challenges that require a community-wide discussion and collaborative efforts to solve them. Several key elements of holdings data are not encoded in MARC or stored in the same ILS module with bibliographic data. This requires a coordinated effort with ILS vendors as well as publishers. In addition, the way that some item related data is organized and managed must be adjusted based on characteristics of each item, e.g., there is no consistency in representing enumeration/volume information for multipart item or serials. In case of items in special collections (and online resources), the eligibility and availability information is hard to capture and represent as LOD without proper semantics. Additionally, gathering this data requires working with systems where the information is stored and updated dynamically.

While the `schema.org` and `BibExtend Community Group`'s recommendations provide libraries a good guideline on how to express holdings data as LOD, it is apparent that further discussion and research are also needed to understand how the library has been creating, using, and managing holdings data, in both data structure and systems. Our analysis and experimentation suggests that libraries should change the traditional way of structuring holdings and item data in the library catalog – from inventory focused to discovery and access focused. Differences between types of data that libraries have and the semantics that `schema.org` has established and the `BibExtend Community Group` recommends should also be reconciled, possibly in conjunction with the vocabularies for holdings data available in the `BibFrame` model led by the Library of Congress.

Finally, it would be helpful for libraries to better understand what today's users use and need for holdings and item related data on the web, to locate and obtain/access the information resources they want. Since not all holdings and item related data are useful for discovery and access services on the web, more research is required on which types of information are beneficial for libraries to expose to the web and to establish the holdings data model in LOD.

References

- BibExtend Community Group. (2014). Holdings via Offer. Retrieved, April 1, 2015, from https://www.w3.org/community/schemabibex/wiki/Holdings_via_Offer.
- Cole, T.W., M.J. Han, W.F. Weathers, and E. Joyner. (2013). Library MARC Records into Linked Open Data: Challenges and Opportunities. *Journal of Library Metadata*, v.13/Issue 2-3: pp. 163-196.
- Library of Congress. (2006). MARC 21 Holdings. Retrieved, April 1, 2015, from <http://www.loc.gov/marc/holdings/hdintro.html>.
- Library of Congress. (2014). MARCXML to MODS 3.5. Retrieved July 9, 2015, from <http://www.loc.gov/standards/mods/v3/MARC21slim2MODS3-5.xsl>.
- Library of Congress. (2015). Bibliographic Framework Initiative: Bibframe. Retrieved, April 8, 2015, from <http://www.loc.gov/bibframe/>.
- Library of Congress. (2015). MARC Standards. Retrieved, April 1, 2015, from <http://www.loc.gov/marc/>.
- International Federation of Library Associations and Institutions. (1998). Functional Requirements for Bibliographic Records: Final Report. Retrieved, April 1, 2015, from <http://archive.ifla.org/VII/s13/frbr/frbr3.htm>.
- OCLC. (2014). OCLC Releases WorldCat Works as Linked Data. Retrieved, April 1, 2015, from <https://www.oclc.org/news/releases/2014/201414dublin.en.html>.

- OCLC. (2015). WorldCat Work Descriptions. Retrieved, April 1, 2015, from <https://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html>.
- schema.org. (2015). What is schema.org? Retrieved, April 8, 2015, from <http://schema.org>.
- Schema Bib Extend Community Group. (2015). Schema Bib Extend Community Group. Retrieved, April 8, 2015, from <https://www.w3.org/community/schemabibex/>.
- Scott, D. (2014). RDFa with schema.org codelab: Library holdings. Retrieved, April 1, 2015, from http://stuff.coffeecode.net/2014/1ld_preconference/rdfa_exercises/2_holdings/.
- Tillet, B. (2004). What is FRBR? A Conceptual Model for the Bibliographic Universe. Library of Congress. Retrieved, April 1, 2015, from <http://www.loc.gov/cds/downloads/FRBR.PDF>.
- Voß, J. and U. Reh. (2015). Document Availability Information API (DAIA). Retrieved July 9, 2015, from <http://gbv.github.io/daiaspec/daia.html>.

Exploratory Analysis of Metadata Edit Events in the UNT Libraries' Digital Collections

Hannah Tarver
University of North Texas
Libraries, USA
hannah.tarver@unt.edu

Mark Phillips
University of North Texas
Libraries, USA
mark.phillips@unt.edu

Abstract

This paper presents the results of an exploratory analysis of edit events performed on records in the University of North Texas Libraries' Digital Collections during calendar year 2014. By comparing the amount of time that editors worked on records for certain item types and collections, we were able to isolate different categories of activities ("creating" vs. "editing") and to generalize rough benchmarks for expected editing durations depending on project criteria.

Keywords: metadata creation; metadata editors; edit events; benchmarks; editing activities

1. Introduction

One ongoing challenge for any metadata creation operation involves estimating the amount of time needed to create (or normalize) metadata for a particular project as well as the costs for doing the work. A reasonable estimate of time helps to build realistic timelines for internal or grant-funded projects, gauge the number of staff needed to meet deadlines, and assess the amount of funding required. To address this need, we decided to perform an exploratory analysis of data within the University of North Texas (UNT) Libraries' Digital Collections.

The Digital Collections comprise three large digital library interfaces: the UNT Digital Library (<http://digital.library.unt.edu>), The Portal to Texas History (<http://texashistory.unt.edu>), and The Gateway to Oklahoma History (<http://gateway.okhistory.org>). The UNT Digital Library primarily contains items owned, licensed, or created by UNT community members. The Portal is collaborative and contains materials owned by more than 250 partner institutions from across the state of Texas, while the Gateway hosts materials owned by the Oklahoma Historical Society. Materials from these collections are in a single, unified infrastructure and all items in our system use the same locally-qualified Dublin Core metadata with twenty-one possible fields (UNT, 2015). Records may be created in-house or by partner institutions, resulting in a large number of editors.

All of the digital library infrastructures for the Digital Collections, including public and administrative interfaces, were built in-house from open source software. Administratively, all item records are accessed via a single metadata editing environment (see Appendix A) locally referred to as the "Edit System." The Edit System loads the current version of a metadata record (which can range from a blank template to a complete record) into a user interface that allows users (i.e., metadata editors) the ability to complete or modify the record and then publish it. At this point, the Edit System saves the most current version and re-indexes the record. Each time an editor interacts with a metadata record, the Edit Event system (see Appendix B) logs the duration and basic metadata information. The analysis presented in this paper is based on events logged by the Edit Event system.

2. Methods

The research questions that guided this exploratory study are: Can metadata event data be used to establish and verify benchmarks within a metadata environment by looking at general

information such as editor or record identity and length of edits? Can metadata edit event data be used to understand the activities of specific users within collections in a metadata system?

Our metadata system creates a log entry when a user opens a record to begin editing, starting a timer for the specific edit session of that record. When the user publishes the record, the Edit Event system queries the log entry and records the duration of the edit in seconds with the editor's username, record identifier, status (hidden or unhidden), record quality -- a completeness metric based on values for eight required fields (title, language, description, subject, collection, partner institution, resource type, format) -- and changes in status or quality (see Table 1). Unless otherwise noted, all duration counts in this analysis are represented in seconds.

TABLE 1: Sample metadata Edit Event system entry.

ID	Event Date	Duration	Username	Record ID	Record Status	Record Status Change	Record Quality	Record Quality Change
73515	2014-01-04T22:57:00	24	mphillips	ark:/67531/metadc265646	1	0	1	0

With this information we can easily see the number of metadata edits on a given day, within the month, and for the entire period we've been collecting data. We can also view the total number of edits, the number of unique records edited, and finally the number of hours that our users have spent editing records within a given period.

We decided to limit this analysis to the calendar year lasting from January 1, 2014 to December 31, 2014, to have a concrete period of time with a reasonable number of data points. The logs contained a total of 94,222 metadata edit events for that year, across 68,758 unique records. These events represent a full range of edit types for materials in our collections. In some cases records were created from blank or near-blank templates by staff members or partner institutions; in other cases, edits were made to correct errors, fix formatting, or add new information to completed records.

In addition to the metadata edit events, we extracted information from the UNT Libraries' Digital Collections related to the individual records: the contributing partner institution, collection code, resource type, and format data for each edited record. We also manually coded the 193 unique metadata editors in the system to classify each as a UNT-Employee or Non-UNT-Employee, and to assign a "rank" of librarian, staff, student, or unknown.

The information was merged and loaded into a Solr index, used as the base datastore for this analysis. We made use of built-in functionality of the Solr index (e.g., StatsComponent, Simple Faceting, and Pivot Faceting) and wrote Python scripts to interact with the data from Solr as needed.

3. Findings

To address the research questions, we first performed basic analysis on the dataset for some of the primary factors including: who is editing the records, what they are editing, and length of edits.

3.1. Who

A total of 193 unique metadata editors logged 94,222 edit events during 2014. As Figure 1 shows, the ten most prolific editors (5% of population) made 57% of overall metadata edits; the graph quickly tapers down to the "long tail" of users who have a lower number of edit events. Since we are reporting on the activities within our own system, it is not surprising that the authors are both listed in the top 5%, as well as others employed in the department.

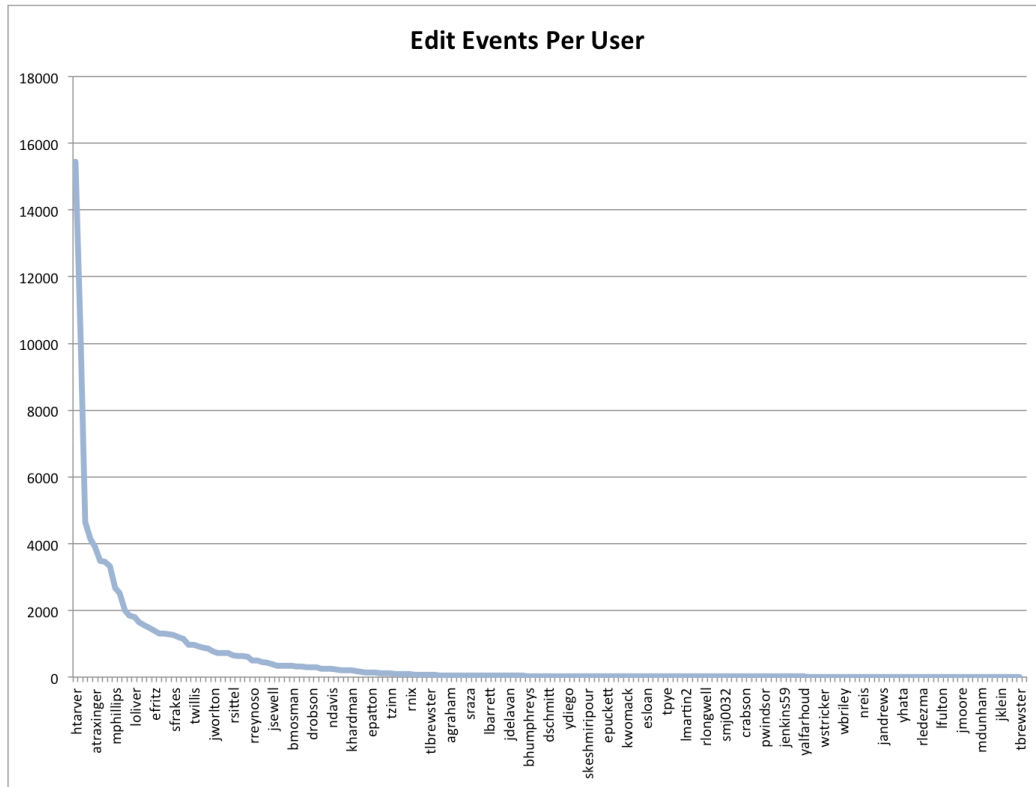


FIG. 1. Distribution of edit events, per editor.

Of the 193 editors in the dataset, 135 (70%) were classified as Non-UNT-Employee and 58 (30%) were classified as UNT-Employee. For the edit events, 75,968 (81%) were completed by a user classified as a UNT employee and 18,254 (19%) by a non-employee user. We also broke this down based on assigned rank of librarian, staff, student, or unknown (see Table 2).

TABLE 2: Statistics for the editors in the system based on their rank.

Rank	Edit Events	Percentage of Total Edits (n=94,222)	Unique Users	Percentage of Total Users (n=193)
Librarian	22,466	24%	16	8%
Staff	12,837	14%	13	7%
Student	41,800	44%	92	48%
Unknown	17,119	18%	72	37%

A clear majority (44%) of all of the edits were completed by students, while librarians and staff members combined accounted for 38% of the edits. The number of students includes both UNT employees -- students employed to do metadata work -- and 65 non-employee students who edited records as part of an assignment in a UNT metadata course.

3.2. What

The dataset contained 94,222 edit events occurring across 68,758 unique records, for an average of 1.37 edits per record. The maximum number of edits for a single record was 45, though most of the records -- 53,213 records (77%) -- were edited just once. Roughly 14% (9,937 records) were edited two times; 5% (3,519 records) were edited three times; records with four or more edits per record only account for 4% of the total dataset.

To see the distribution of edits, we categorized records by the partner institution listed in each record and analyzed statistics for the ten most represented partners in the dataset (see Table 3).

TABLE 3: Most edited items by partner institution.

Partner Code	Partner Name	Edit Count	Unique Records Edited	Unique Collections
UNTGD	UNT Libraries Government Documents Department	21,932	14,096	27
OKHS	Oklahoma Historical Society	10,377	8,801	34
UNTA	UNT Libraries Special Collections	9,481	6,027	25
UNT	UNT Libraries	7,102	5,274	27
PCJB	Private Collection of Jim Bell	5,504	5,322	1
HMRC	Houston Metropolitan Research Center at Houston Public Library	5,396	2,125	5
HPUL	Howard Payne University Library	4,531	4,518	4
UNTCVA	UNT College of Visual Arts and Design	4,296	3,464	5
HSUL	Hardin-Simmons University Library	2,765	2,593	6
HIGPL	Higgins Public Library	1,935	1,130	3

Many partners who are heavily represented have edits spread across multiple collections. However, there are also differing trends regarding the ratio of edits to records. Figure 2 quickly shows which partners often make multiple edits per record as opposed to those partners that tend to have only one record edit event per record. In some cases, such as the editing done by Houston Public Library, the number of edits is roughly double the number of records, versus editing relationships that are nearly one-to-one (e.g., Hardin-Simmons University Library).

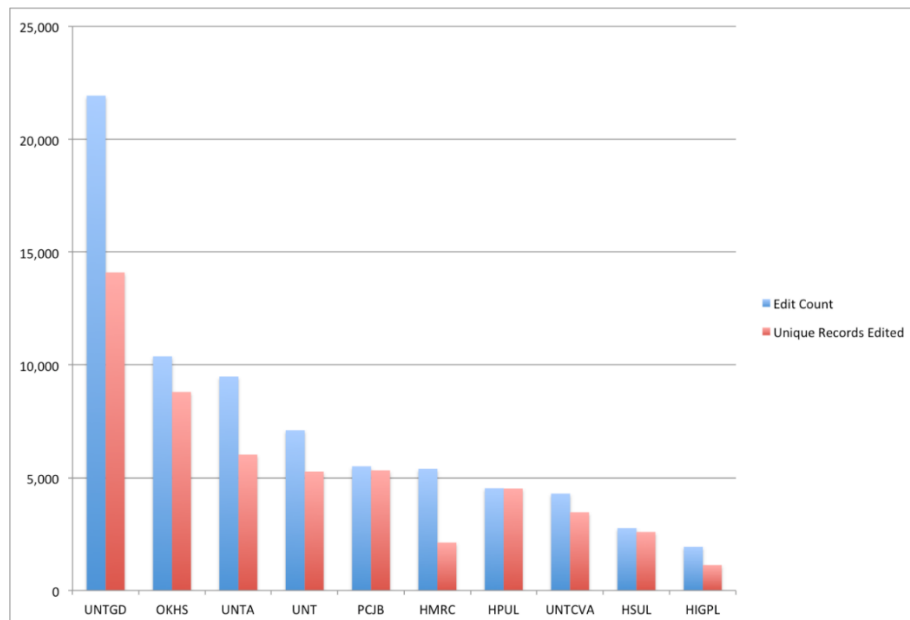


FIG. 2. Comparison between edit count and unique records for the top ten partners.

Since edits may be distributed across multiple collections (see Table 3), we analyzed edit events by collections and determined the ten collections that had the most edited items (see Table 4).

TABLE 4: Most edited items by collection.

Collection Code	Collection Name	Edit Events	Unique Records Edited
TLRA	Texas Laws and Resolutions Archive	8,629	5,187
ABCM	Abilene Library Consortium	8,481	8,060
TDNP	Texas Digital Newspaper Program	7,618	6,305
TXPT	Texas Patents	7,394	4,636
OKPCP	Oklahoma Publishing Company Photography Collection	5,799	4,729
JBPC	Jim Bell Texas Architecture Photograph Collection	5,504	5,322
TCO	Texas Cultures Online	5,490	2,208
JJHP	John J. Herrera Papers	5,194	1,996
UNTETD	UNT Theses and Dissertations	4,981	3,704
UNTPC	University Photography Collection	4,509	3,232

The distribution of edit events and unique records also varies by collection. Since items can have assignments to multiple collections, some of the data in Table 4 overlaps. For example, the John J. Herrera Papers were part of the Texas Cultures Online project, which explains why the editing trends look similar (see Fig. 3). However, there were other edits to the Texas Cultures Online project which were not part of the Herrera papers, so the numbers are not an exact match.

Figure 3 shows the relation of edit events and unique records by collection. There are some slightly different trends, but this information is helpful in our system because collections often encompass discrete projects, while edits to partner items may be spread across multiple projects.

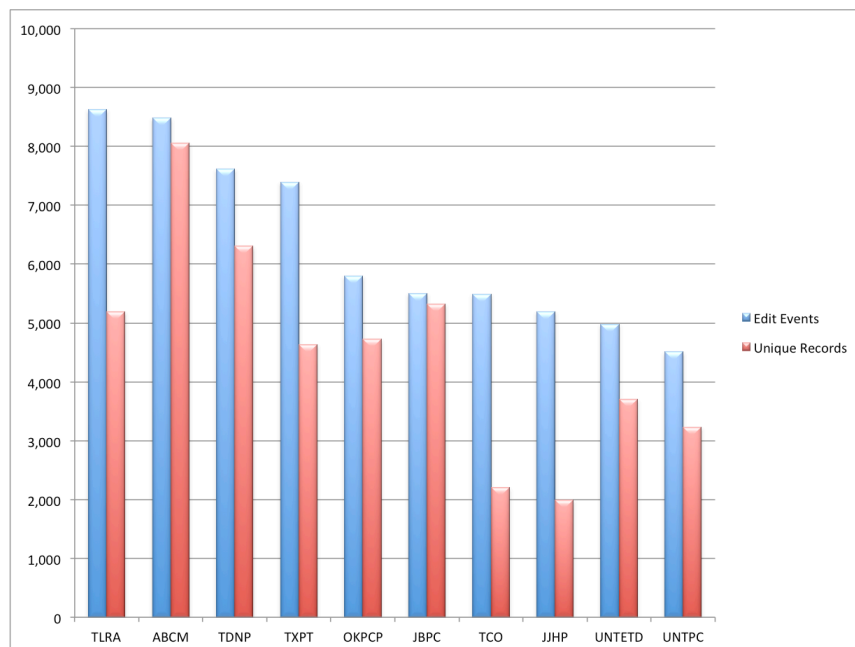


FIG. 3. Comparison between edit count and unique records for the top ten collections.

3.3. How Long

Without a time aspect, it would be difficult to formulate benchmarks or generalize conclusions from the raw data. The duration of edits in this dataset ranged from only 2 seconds to over 119 hours. To better visualize the distribution, the duration of each edit event was grouped into “buckets” of hours and minutes. A majority of edit events -- 93,378 (99%) -- lasted for 60

minutes or less. Of these events that happened within an hour, 75,981 (81%) of the events lasted less than 6 minutes and 17,397 (19%) lasted 7-60 minutes (see Fig. 4).

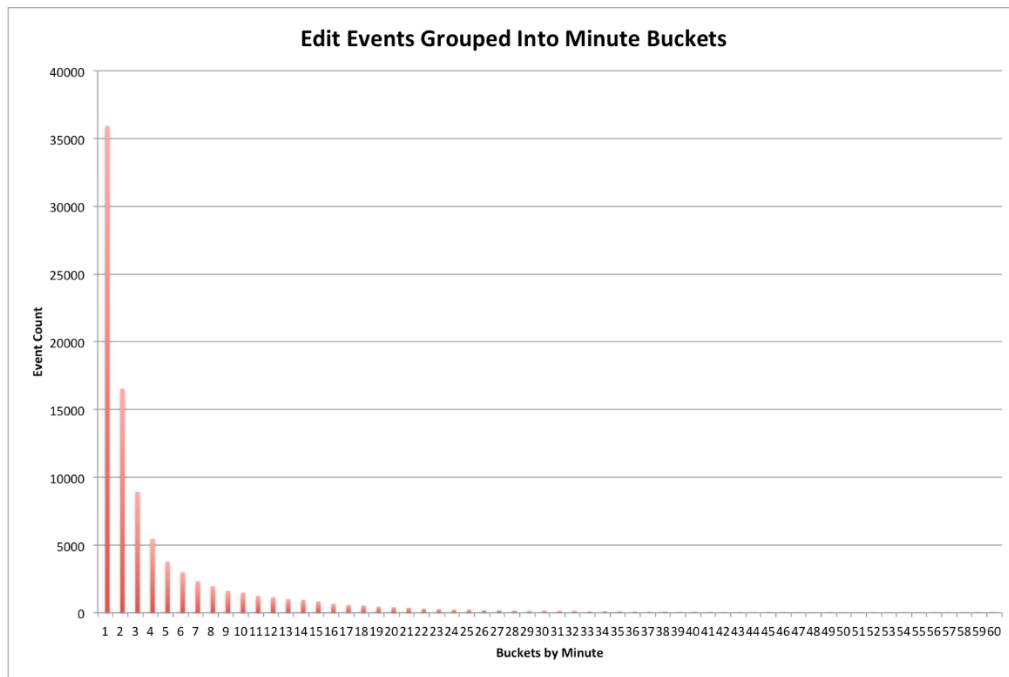


FIG. 4. Distribution of edits up to 60 minutes in duration (n-93,378)

Since a relatively large number of events (35,935) lasted less than one minute, we graphed this subset to see where those edit events fell within the distribution by number of seconds (see Fig. 5).

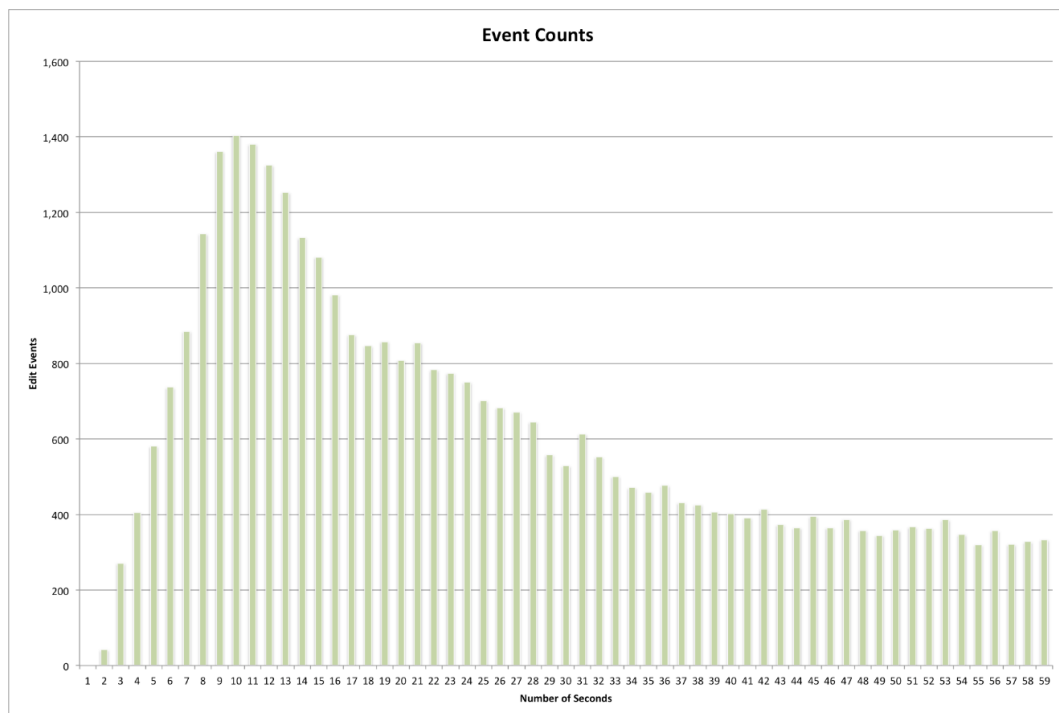


FIG. 5. Distribution of edits up to 60 seconds in duration (n-35,935)

We wanted to eliminate excessively long edits that do not represent “normal” editing and may be errors. Based on the distributions, we chose a range encompassing a majority (97.6%) of edits, establishing a ceiling of 2100 seconds (35 minutes). This threshold leaves 91,916 remaining events; the other 2,306 events were ignored for all further calculations. The average duration of edits lasting 2100 seconds or less was 233 seconds with a standard deviation of 345.48.

4. Discussion

Based on these who/what/how long questions, we wanted to draw reasonable benchmarks for editing activities in our system and to better gauge editing activities. Further comparisons show where times based on kinds of records and edits can provide useful information.

4.1. Time by Item Type

Some records may take longer than others because more information is available to enter, or because it takes more time to skim information on text items than to look at an image and describe it. Although there will always be outliers, the average amount of time by resource type should demonstrate general trends. Table 5 displays edit times by type, including minimum and maximum duration, number of records, total edit time, average (mean), and standard deviation (stddev).

TABLE 5: Average duration of edits (in seconds) by resource type.

Resource Type	Min	Max	Records	Total Time	Mean	Stddev
image_photo	2	2,100	30,954	7,840,071	253.28	356.43
text_newspaper	2	2,084	11,546	1,600,474	138.62	207.30
text_leg	3	2,097	8,604	1,050,103	122.05	172.75
text_patent	2	2,099	6,955	3,747,631	538.84	466.25
physical-object	2	2,098	5,479	1,102,678	201.26	326.21
text_etd	5	2,098	4,713	1,603,938	340.32	474.40
text	3	2,099	4,196	1,086,765	259.00	349.67
text_letter	4	2,095	4,106	1,118,568	272.42	326.09
image_map	3	2,034	3,480	673,707	193.59	354.19
text_report	3	1,814	3,339	465,168	139.31	145.96

As expected, text items tend to take longer, though edit time for photographs is also high. This may be due to the number of photograph records created from scratch, especially when other sources were consulted. The largest spike is in the average time for patent records; this is likely because patent records are being created from near-blank templates and require a large amount of information. We also use patent records for library students or volunteers to experiment with creating metadata, so a number of these editors are new and may tend to take longer than experienced editors.

Based on this information, we can say that editors should expect to spend roughly 10 minutes per patent record, once they are familiar with the system. Some item types are more ambiguous. For example, photographs have a lower average time, but they are a mix of records written from scratch and those edited less extensively. It is still helpful for editors and supervisors to know that most of the time, editing photograph records for longer than 5 minutes is excessive. In this case, more information about the collection would provide a better sense of expected average times.

4.2. Time by Collection

In general, we have internal knowledge about which collection records were primarily created from scratch versus those that required cleanup or less extensive additions. While editors may

conduct different kinds of activities within a collection, the average amount of time (see Table 6) should still give a sense of time spent on records, especially if combined with other information.

TABLE 6: Average duration of edits (in seconds) by collection

Collection Code	Collection Name	Min	Max	Edit Events	Duration Sum	Mean	Stddev
TLRA	Texas Laws and Resolutions Archive	2	2,083	8,418	1,358,606	161.39	240.36
ABCM	Abilene Library Consortium	3	2,100	5,335	2,576,696	482.98	460.03
TDNP	Texas Digital Newspaper Program	3	2,095	4,940	1,358,375	274.97	346.46
TXPT	Texas Patents	5	2,084	3,946	563,769	142.87	243.83
OKPCP	Oklahoma Publishing Company Photography Collection	4	2,098	5,692	869,276	152.72	280.99
JBPC	Jim Bell Texas Architecture Photograph Collection	3	2,095	5,221	1,406,347	269.36	343.87
TCO	Texas Cultures Online	2	1,989	7,614	1,036,850	136.18	185.41
JJHP	John J. Herrera Papers	3	2,097	8,600	1,050,034	122.10	172.78
UNTETD	UNT Theses and Dissertations	2	2,099	6,869	3,740,287	544.52	466.05
UNTPC	University Photography Collection	3	1,814	2,724	428,628	157.35	142.94

Table 6 presents average edit times by collection for the ten most edited collections. The same spike for patents appears here since the Texas Patent collection has a one-to-one relationship with the patents (resource type). There is also a higher-than-expected average for the Jim Bell Texas Architecture Photograph Collection, even when compared to similar collections (e.g., the University Photography Collection). However, the primary editor for this collection often opened many records so that they would be loaded and waiting; this action skewed the data since the system calculates duration based on when the record was opened, rather than on activity.

4.3. User Activities

Our second research question focused on identifying kinds of editing activities by user. We looked at statistics for the ten most active editors (see Table 7).

TABLE 7: Statistics of edits by user for the top ten editors.

Username	Min	Max	Edit Events	Duration Sum	Mean	Stddev
htarver	2	2,083	15,346	1,550,926	101.06	132.59
aseitsinger	3	2,100	9,750	3,920,789	402.13	437.38
twarner	5	2,068	4,627	184,784	39.94	107.54
mjohnston	3	1,909	4,143	562,789	135.84	119.14
atraxinger	3	2,099	3,833	1,192,911	311.22	323.02
sfisher	5	2,084	3,434	468,951	136.56	241.99
cwilliams	4	2,095	3,254	851,369	261.64	340.47
thuang	4	2,099	3,010	770,836	256.09	397.57
mphillips	3	888	2,669	57,043	21.37	41.32
sdillard	3	2,052	2,516	1,599,329	635.66	388.30

In some cases, editing activities are more apparent when editors are working at different levels on a set of items. Figure 6 shows the average edit by editor for legislative text (type) items.

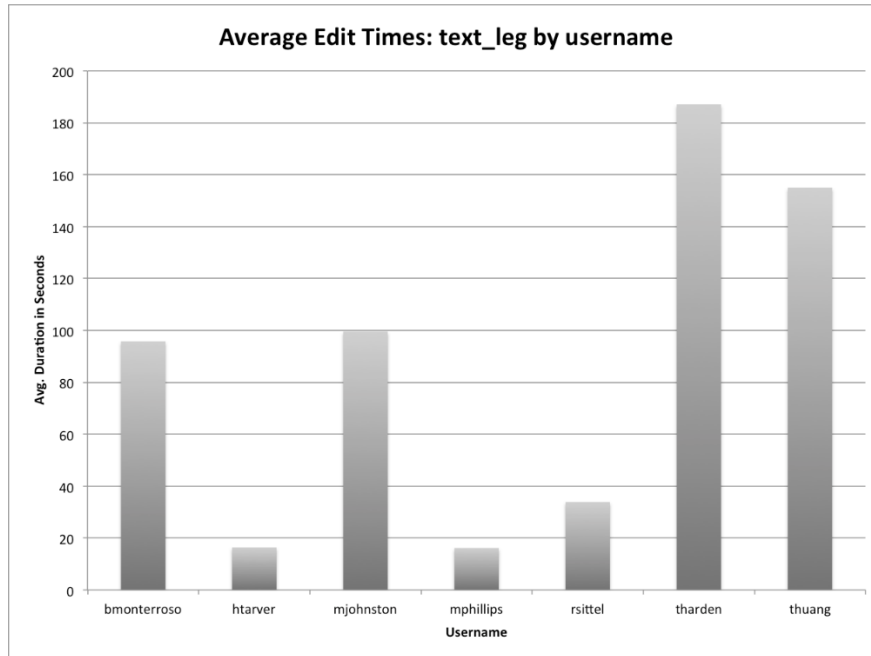


FIG. 6. Average duration of edits by user for legislative text items.

Most records with a legislative text resource type had little starting information. Several of the editors in the set -- htarver, mphillips, and rsittel (all classified as librarians) -- have significantly lower average times than other editors, suggesting that they were performing less extensive edits, compared to editors who spent longer amounts of time on the items. It also shows a trend where students (mjohnston, tharden, and thuang) are primarily “creating” records by adding significant amounts of information while librarians and staff (including bmonterroso) generally supervise by making minor changes and corrections.

While this is not entirely conclusive, we can distinguish “new record creation” versus “minor edits” when compared to average times of similar types, other items in the collection, or against various users. In the future, this provides an opportunity to isolate activities and find a reasonable average time per record creation based on comparable collections.

5. Conclusions

This paper describes an exploratory analysis of the 94,222 metadata edit events logged by 193 editors in the UNT Libraries’ Digital Collections from January 1 to December 31, 2014. Based on data collected from the Edit Events system and information known about the records and editors, we discovered that multiple variables affect editing times, but we can generalize about the kinds of activities and how close various edits come to a “normal” or average duration.

5.1. Benchmarks

We particularly wanted to know if we could use gathered editing data to define general characteristics for certain kinds of editing projects within our system. Overall, we discovered that edits of any kind are unlikely to take more than 35 minutes and the average time for those edits is only three minutes and fifty-four seconds.

When monitoring a project, it may be useful to see if the average time is near four minutes or if it differs significantly. However, based on the analyses in the previous section, we can also take into account the resource type and kind of collection. For a text-based collection, we would expect the average time for “creation” to be closer to ten minutes, rather than the system average.

Additionally, we could use average duration to determine the kinds of edits -- in particular, whether users are acting as “creators” and primarily making large additions or significant changes versus “editors” enacting relatively minor edits and corrections. Likewise it should be possible to identify users of browser automation tools, such as Selenium¹, to streamline the editing process.

Distinguishing between “creators” and “editors” could be applied to tracking projects when users with different roles are working on a collection; e.g., two editors “creating” records and keeping them hidden while a third (supervisor) reviews and publishes the records. We would expect the first two editors to have similar duration averages while the third user might have a substantially lower average. Project-level benchmarks could be based on average times by role.

In terms of our research questions, we *can* determine the general kinds of editing activities and create project-based benchmarks based on similar project variables from information in this study.

5.2. Next Steps

Building on this initial study, several comparisons could augment precision in our benchmarks. Pairing the number of metadata edits per collection and partner institution with the average user durations would make it possible to identify administrative editors in the system, or those who are metadata “creators.” This may lead to more accurate item-type or collection-level benchmarks when general averages do not fit a project well.

Additionally, it is possible to calculate the total amount of time spent on a given record by adding the edit durations, either by one user or for all users. This information could be valuable for establishing the average amount of time needed to fully complete a metadata record.

One area of interest, which we were not able to explore in this particular study, is to assign hourly costs to users based on ranks (librarian, staff, student, or unknown) and to calculate approximate costs paid by UNT for employee editors versus time “donated” by non-employees. Additionally, if more information can be gathered about the editors -- such as metadata experience -- it may be possible to determine if other variables affect average durations and the cost-per-record.

5.3. Further Study

The analysis presented in this paper is a first step. Although statistics for other institutions may be affected by differences in system interfaces or kinds of collections, staff at other repositories could collect similar data to see if trends match our findings and build benchmarks for editing their collections. It may also be helpful for other groups to use similar criteria identified in this study as a starting point, particularly resource type and collection information since those seem to provide a reasonable cross-section for benchmarking average metadata creation times for many materials. Additional work could also pinpoint which criteria or combination of criteria are most useful for outlining benchmarks based on this kind of data.

Exploring data and the aggregate statistics in this dataset may allow researchers to help metadata editors and administrators produce higher quality metadata records for less overall cost.

References

- Phillips, Mark Edward. (February 2015). UNT Libraries 2014 Editors Event Dataset [dataset]. <http://digital.library.unt.edu/ark:/67531/metadc502990>
- University of North Texas Libraries (2015). Input Guidelines for Descriptive Metadata (revised version). Retrieved from <http://www.library.unt.edu/digital-projects-unit/input-guidelines-descriptive-metadata>

¹ <http://www.seleniumhq.org/projects/ide/>

Appendix A: UNT Editing System

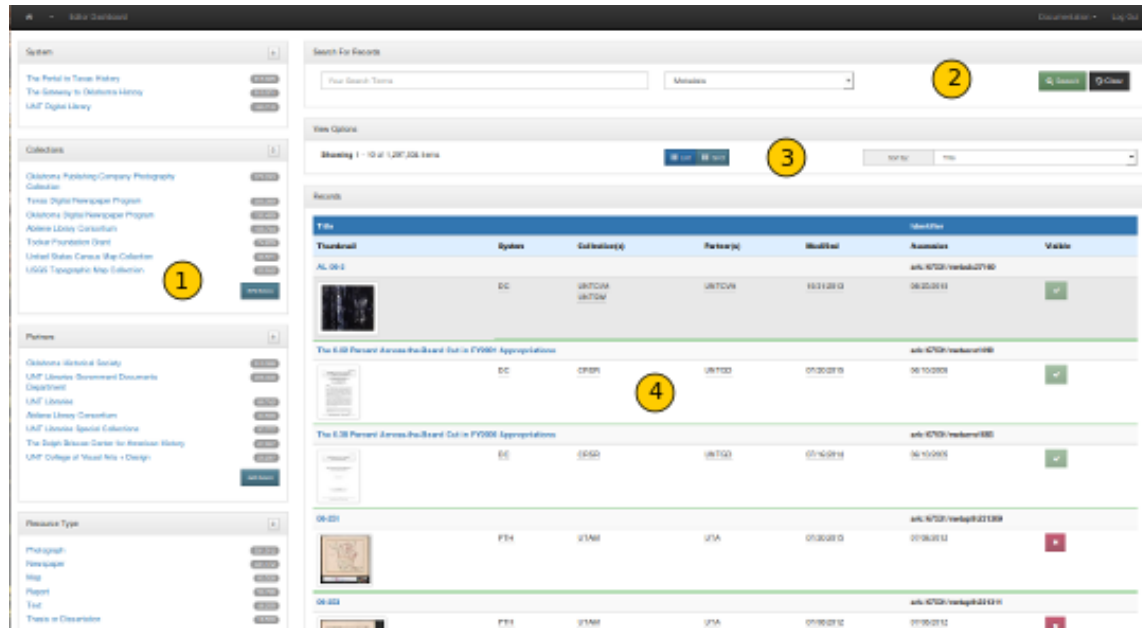


FIG. 7. Screenshot of the user Dashboard in the Editing System.

When a user logs into the Edit System, he sees a list of all records for which he has access. Clicking a title or thumbnail will open the item record in a new tab or window (see Fig. 8).

Dashboard Features

1. Facet options to narrow records by system interface, collection, partner, resource type, and public visibility (when applicable).
2. Search bar to find terms in records, with a drop-down menu to limit searches to a specific field.
3. Options to display item records in a thumbnail grid or list (shown here) and a drop-down menu to sort by the dates records were added or modified, item creation dates, or unique ARK identifiers.
4. List of item records displaying the title, thumbnail, system, partner, collection, date added and modified, ARK identifier, and public visibility status for each.

FIG. 8. Screenshot of an item record in the Editing System.

This is the view of a metadata record containing an incomplete template for an item.

Record Fields

1. Text box(es) and/or drop-down menu(s) appropriate for the field are displayed in a bounded box with a title bar.
2. The title bar for each field includes a “Help” link to the guidelines for the field (which open in a pop-up modal), as well as an icon to collapse the field.
3. At the bottom of the field, buttons allow a user to insert symbols and add or remove field entries.

Navigation

4. All of the fields are listed on the right side of the screen and are clickable so that an editor can go directly to a specific field. A bubble next to each field title lists the number of entries in the field; the bubbles are color-coded to show if required fields have values (red = no value, green = value present) and to highlight invalid dates or insufficient subjects (yellow).
5. Clicking the thumbnail opens a new tab displaying all images (pages, views, etc.) for the item.
6. Radio buttons let an editor change the status (visible to or hidden from the public) and the “Publish” button saves all changes to the record.

Appendix B: UNT Edit Event System

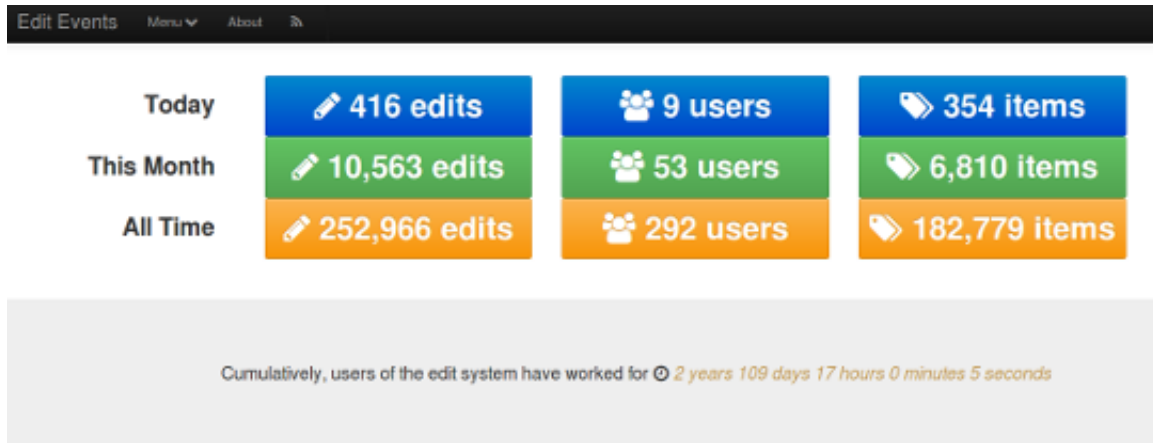


FIG. 9. Screenshot of the Edit Event System dashboard.

The Edit Event system dashboard displays current statistics at the time the page is accessed. Each of the buttons is clickable, to show additional statistics for specific dates, users, or records.

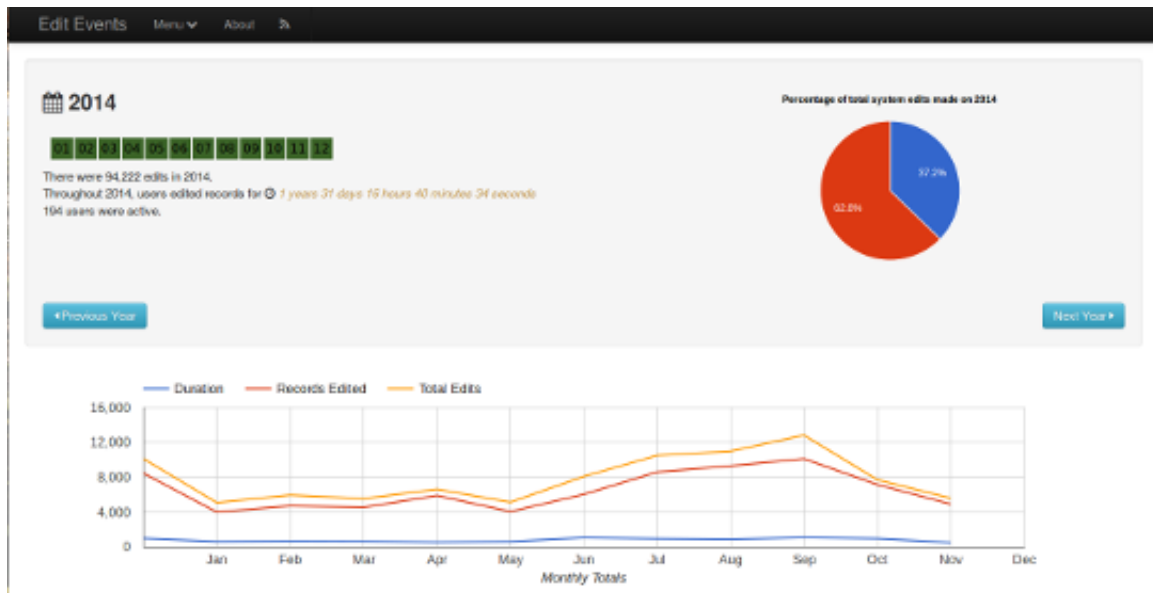


FIG. 10. Screenshot of the statistics page for the 2014 calendar year.

This is an example of a more specific page, showing overall information for 2014. Months are listed across the top to limit by a particular month, followed by a summary of various statistics associated with the chosen date (e.g., total edits, total time, number of editors, etc.). The pie chart shows how the number of edits during the year (blue, 37.1%) compares to the number of other edits (red, 62.9%) logged by the system. The graph at the bottom has lines showing average duration, number of records edited, and total edits throughout the year. Similar statistics are displayed at every level, depending on relevant information for that date, user, or record.



Application Profiles: Reflections & Developments—Session 2

Evolution of an Application Profile: Advancing Metadata Best Practices through the Dryad Data Repository

Edward M. Krause
Metadata Research Center
Drexel University, USA
edward.krause@drexel.edu

Erin Clary
Metadata Research Center
Drexel University, USA
eclary@datadryad.org

Adrian Ogletree
Metadata Research Center
Drexel University, USA
aogletree@drexel.edu

Jane Greenberg
Metadata Research Center
Drexel University, USA
janeg@drexel.edu

Abstract

Dryad is a general-purpose curated repository for data underlying scholarly publications. Dryad's metadata framework is supported by a Dublin Core Application Profile (DCAP, hereafter referred to as application profile). This paper examines the evolution of Dryad's application profile, which has been revised over time, in an operational system, serving day-to-day needs of stakeholders. We model the relationships between data packages and data files over time, from its initial implementation in 2007 to its current practice, version 3.2, and present a crosswalk analysis. Results covering versions 1.0 to 3.0 show an increase in the number of metadata elements used to describe Dryad's data objects in Dryad. Results also confirm that Version 3.0, which envisioned separate metadata element sets for data package, data files, and publication metadata, was never fully realized due to constraints in Dryad system architecture. Version 3.1 subsequently reduced the number of metadata elements captured by recombining the publication and data package element sets. This paper documents a real world application profile implemented in an operational system, noting practical system and infrastructure constraints. Finally, the analysis presented informs an ongoing effort to update the application profile to support Dryad's diverse and expanding community of stakeholders.

Keywords: metadata; metadata schema; application profile; DCAP; Dryad.

1. Introduction

Dryad has been supported by a metadata application profile from its launch in 2007 through the present day (Dryad Data, 2015). An application profile "consist[s] of data elements drawn from one or more namespace schemas combined together by implementors and optimised for a particular local application" (Heery & Patel, 2000). A data element refers to a metadata field, and a namespace schema, or a metadata schema, is a set of standardized metadata elements.

The application profile approach was endorsed by Dryad team members from the beginning, given the need for Dryad metadata to interoperate with other data efforts, and given the desire of Dryad's metadata R&D team to align with semantic web developments and to keep current with metadata developments. Application profiles promote data sharing, interoperability, and linked data, which are all central to the overarching mission of Dryad.

Dryad has been operational since 2008, and has grown at a fairly rapid pace, expanding to accommodate more disciplines and stakeholder organizations. This growth has had an impact on Dryad's functional requirement and day-to-day workflows, expanding the menu of options. These changes have had a significant impact on Dryad's metadata application profile. In this paper, we perform a crosswalk analysis, present domain models, and evaluate individual metadata elements and refinements that have changed over time. The paper also serves to document the change in an

application profile over time and to produce an updated representation of Dryad's current metadata practice.

1.1. What is Dryad?

Dryad is a curated digital repository for data underlying peer-reviewed scholarly literature. The stated mission of the repository is to "make the data underlying scholarly publications discoverable, accessible, understandable, freely reusable, and citable" (Dryad, 2015). Dryad is also committed to the long-term preservation of archived data (Mannheimer et al., 2014). While Dryad began as an infrastructure for data archiving in evolutionary biology and ecology, the scope of the repository has since expanded. Dryad has developed into a general-purpose repository for long-tail scientific data, and the repository currently accepts data from a wide variety of disciplines, including medical and social sciences.

Each data package in Dryad is linked to its associated publication, and Dryad stores metadata related to the data package and its files, in addition to metadata derived from the publication. Dryad works with a data package model, in which a data package can have one or more data files. Dryad's chief mission is to make data discoverable and reusable for scientific endeavors. Metadata is essential for these steps, and for fulfilling Dryad's mission.

1.2. Dryad's Early Application Profile Work (2007-2009)

Since its origins, Dryad has actively incorporated the Dublin Core Abstract Model (DCMI, 2007), adhering to the Singapore Framework for Dublin Core Application Profiles (DCMI, 2008), into a metadata best practice (Powell et al., 2007; Nilsson et al., 2008; Greenberg et al., 2009). These two abstract information models, developed by the Dublin Core community, represent efforts to move from the resource-driven legacy approach representing an information package toward focusing on the component parts of a resource description. The initial goals of developing an application profile for Dryad were twofold; an immediate short-term concern was to make content available in DSpace through an XML schema, and in the long-term, to align with the Semantic Web (Greenberg et al., 2009).

The first version of Dryad's application profile (v1.0) was developed in 2007, before the release of the Singapore Framework guidelines. Although the Singapore Framework had not yet been published, development of Dryad's metadata application profile still began with the critical first steps of defining the repository's functional requirements and creating a domain model, as prescribed in the Guidelines for Dublin Core Application Profiles (Coyle & Baker, 2009). These first steps are reported on in more detail in Dube et al. (2007) and White et al. (2008).

TABLE 1: Dryad DCAP v.3.1: Metadata elements (Dryad, 2013).

Data Package	Data File
dcterms:type	dcterms:type
dcterms:creator	dcterms:creator
dcterms:dateSubmitted	dcterms:dateSubmitted
dcterms:available	dcterms:available
dcterms:title	dcterms:title
dcterms:identifier	dcterms:identifier
dcterms:description	dcterms:description
dcterms:subject	dcterms:subject
dwc:scientificName	dwc:scientificName
dcterms:spatial	dcterms:spatial
dcterms:temporal	dcterms:temporal
external	embargoedUntil
dcterms:references	dcterms:rights
bibo:pmid	dcterms:format
bibo:Journal	dcterms:provenance
dcterms:hasPart	dcterms:isPartOf

1.3. Dryad, DSpace, and Further Application Profile Development (2013-2015)

The most current version of the application profile is v3.1, published in an XSD file (Dryad, 2013). The metadata elements included in v3.1 are listed in Table 1. Elements that are shaded green are the ones that are used to describe both the data package and file. These elements are intended to document bibliographic metadata of the associated publication, scope and coverage of the data files, and the relationship between the data files, the data package, and the publication, each of which is represented by a unique identifier. The profile includes elements from several namespaces, including Dublin Core (DCMI, 2012), Darwin Core (Darwin, 2015), and Dryad's own namespace. Dryad has been implemented on version 1.8 of the DSpace framework (DSpace, 2015). While the latest version of DSpace released, as of this publication, is version 5.0, Dryad has not upgraded to a later version of the framework due to the risk of unforeseen upgrade incompatibilities with the extensive customizations of the system architecture made by Dryad developers. Though Dublin Core does not support dot-notation for representing metadata elements and the associated refinements (e.g. `dcterms:coverage.spatial`), DSpace continues to use this type of notation internally to represent metadata elements. During automated metadata harvesting, internal metadata elements are converted to Dublin Core compliant properties from the *terms* namespace.

Dryad is built on an early version of DSpace and elements are stored internally. DSpace is among one of the most popular repository software used for digital libraries, storage of offprints, and other digital creative outputs of an institution. Among several well-known DSpace users are Cornell University Libraries, Deep Blue at the University of Michigan, and Rice University's TIMEA digital archive. DSpace was selected for Dryad because of its open source status, its user-friendly interface for scientists/researchers as depositors, and because it could be installed out of the box. Dryad has worked with Atmire (<http://atmire.com/website/>) since the beginning to better accommodate scientific data deposits.

Ongoing development of an operational system, with real users and day-to-day needs, has been an exciting undertaking for the Dryad team. The progress has been consistent, keeping Dryad fully functional, although, as one may anticipate, there have been delays in keeping pace with the most current DSpace release, particularly given the unique nature of Dryad. Another important point is that DSpace provides access to an extensive list of Dublin Core metadata properties along with properties from additional namespaces within the curation module; however, the current metadata infrastructure doesn't fully align with the DCMI's DCAP for rendering RDF metadata, and the syntactic encoding differs. Metadata generated via DSpace can be converted to RDF, although this has not been a chief priority for Dryad at this time, with current day-to-day, real-world needs servicing clients and making descriptions accessible. The aim of being fully compliant with DCMI, aligning with the Singapore Framework, and the DCAM (Dublin Core Abstract Model) is part of Dryad's two-pronged approach, and has been documented in Greenberg et al., (2009). This paper presents an account of the activity that is impacting the day-to-day work, and the guiding research objectives are outlined in the next section.

2. Research Objectives

This study is the first step in a larger process to document and assess Dryad's metadata application profile. Dryad's initial metadata scheme was devised to allow for data ingest, and to support preservation, access, and basic usage of data (Dube, 2007). Dube et al. (2007) also proposed long-term goals for the metadata scheme, including expanded support for data use, extended interoperability and support for semantic web functionalities.

Dryad's initial disciplinary focus was evolutionary biology. Today, the repository is still heavily in the bioscience area, although Dryad is promoted as a general-purpose repository, and there is a growing representation from a wide array of disciplinary fields, ranging from the biomedical field to physics, chemistry, information science, and social sciences. This change, and stakeholder growth (including more publishers and organizations) has resulted in new functional requirement, which in turn have had an impact on the application profile. Given the pace of

change, it seems timely to revisit the application profile work and document the current practice. The goal of the research reported on in this paper is to examine how Dryad's application profile has evolved from its first inception in 2007 as version 1.0 through the last update in 2013 as version 3.1. This study will document changes in the element set over time. An end goal of this study is to align Dryad's application profile with current practice as version 3.2 and to propose next steps to update the application profile. This will help Dryad to maintain high quality metadata practice, and help provide a platform for attaining higher-level objectives of automatic data synthesis as described in 2007.

3. Methods

To investigate the goals and methods outlined in Section 2, we used a crosswalk analysis to compare each version of the application profile and modeled the relationship between data package, data file and publication that was represented by each application profile. While crosswalk analyses are primarily used to facilitate interoperability among applications that may use different metadata schemas by mapping metadata elements, semantics, and syntax from each schema to determine their compatibility (NISO, 2004), we conducted a modified crosswalk analysis to examine changes in metadata usage across the different versions of Dryad's application profile. Domain models define the basic structures and relationships of digital entities (Nilsson et al., 2009). In Dryad, each entity - data package, data file and publication - is described by a set of metadata elements. Changes in the domain models across application profile versions reflect changes identified in the crosswalk analysis. Each version of the application profile was compared to the previous iteration, and changes in element usage were documented. Lastly, an updated version of the application profile, version 3.2, was created to report on current metadata practices in Dryad.

4. Results and Discussion

The results and contextual discussion that follow detail the crosswalk analysis, Dryad's changing domain models, and version changes.

4.1. Crosswalk Analysis

The Dryad application profile has drawn from multiple metadata schemas throughout its version history. The current profile includes elements from Dublin Core (namespace: dcterms), Darwin Core (namespace: dwc), and Publishing Requirements for Industry Standard Metadata (namespace: prism) (Idealliance, 2015). The application profile also includes Dryad namespace elements, which represent concepts required for repository functionality that were not found in other schemas. For instance, Dryad captures the number of page views and downloads of each data file with the elements `dryad:pageviews` and `dryad:downloads`. As mentioned earlier, DSpace uses a dot-notation to express elements and their refinements internally, and this is how some metadata elements will be described in the Results and Discussion. Table 2 explains the relationship between Dryad/DSpace internal elements and their corresponding external notations as they are represented in automated metadata harvests.

Early versions of the application profile included elements from Data Documentation Initiative (namespace: DDI) (DDI, 2009), Journal Publishing Tag Set (namespace: journalpublishing3) (NCBI, 2012), Preservation Metadata: Implementation Strategies (namespace: PREMIS) (LoC, 2015), and Bibliographic Ontology Specification (namespace: bibo) (Bibliographic, 2009); however, elements from these schemas are not currently used. Many of the metadata elements from the discontinued schemas are now represented as Dublin Core refinements. For instance, version 2.0 used elements from the PRISM and Journal Publishing Tag Set schemas to store publication citation metadata, while version 3.0 replaced and expanded upon the PRISM concepts with elements from the Bibliographic Ontology Specification. In versions 3.1 and 3.2, the elements used to store citation information were collapsed into a single field, `dcterms:identifier`.

The crosswalk analysis revealed four possible cases for each metadata element in the application profile: 1) The element and the concept it represents (an element-concept pair) did not change, and is present in all iterations of the application profile. 2) The concept did not change, but the element that was used to represent that concept did change from version to version. 3) Elements and concepts are added, and 4) Elements and concepts are phased out.

Metadata elements that are used in each version of the application profile include those that represent descriptive, spatial and temporal characteristics, digital identifiers, types, relationships, subjects, and taxonomic classification. Other metadata concepts have remained constant through each version of the application profile, but are represented by different metadata elements over time. For instance, the embargo end date, which is the date on which a data file will be made available for download, was initially recorded at `dcterms:available`. This concept was later represented by the element `dcterms:embargoedUntil`, while `dcterms:available` was repurposed to represent the date and time a curator approved a data package into the archive. This definition of `dcterms:available` was more congruent with the Dublin Core definition of this term as a “date (often a range) that the resource became or will become available” (DCMI, 2012). However, the metadata describing a data file may be made available at the public website before the file itself is available for download, hence the embargo date refinement for data files within a data package.

Each version of the application profile is a snapshot of Dryad’s workflow and functionality at a particular point in time. While many of the elements of versions 1.0 and 2.0 were phased out prior to the current version, version 3.0 introduced multiple concepts and elements that are currently used; these element-concept pairs chronicle the evolution of repository functionality. For instance, an element to record provenance metadata, `dcterms:description.provenance`, was added in version 3.0. Metadata, including date, time and name of the person who performed an action, are automatically captured at ingest, and each time a data package changes workflow stages. The crosswalk analysis also depicts a more recent increase in the number of concepts and elements added to the application profile in version 3.2. For instance, publication blackout dates allow for automated release of submissions to the archive, correlating to the expected release of the article online by the publisher. Recent element changes demonstrate an increase in advanced functions, including automation of certain curation tasks.

4.2. Dryad’s Changing Domain Models

Comparison of the domain model versions (Figure 1) provides additional context to the application profile version changes.

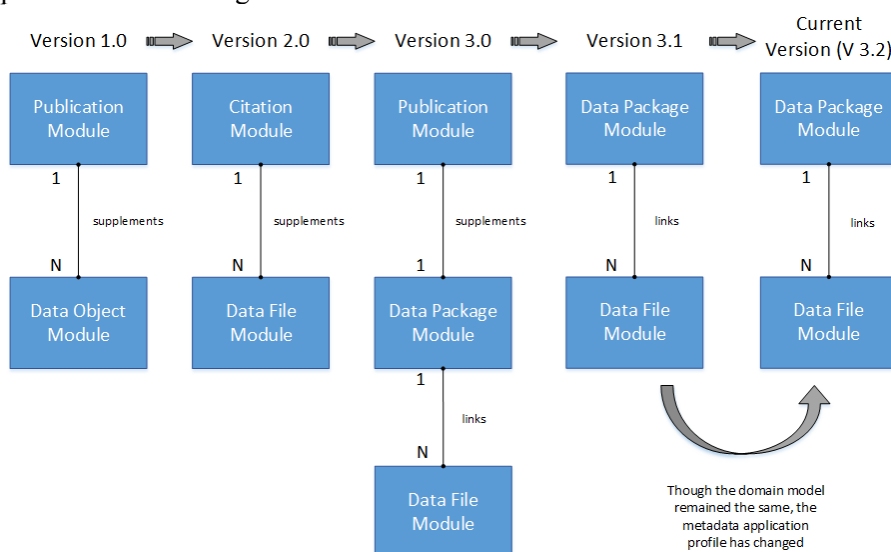


FIG. 1: Dryad domain model versions.

Versions 1.0 and 2.0 use a similar domain model, with a single publication or citation module linked to multiple data objects or files. The publication/citation modules contain metadata pertaining to the publication associated with the archived datasets as well as metadata that links to the data file modules. In versions 1.0 and 2.0, one or more data files could be associated with one and only one publication or citation. Elements pertaining to the associated journal citation were added in version 2.0, including article title, journal name, volume and issue, which increased the granularity with which journal citations were captured in Dryad.

The evolution from version 2.0 to 3.0 of the application profile domain model shows an expanded set of entities, where the publication module is split from the data package module and the data package module is linked to the data file module. Additional journal citation elements were added in version 3.0, increasing the granularity of the journal citation concept. The additional journal citation elements include ISSN/EISSN, PMID, and status. At this point in the application profile development, provenance metadata was also included to track workflow step changes and the users who perform the workflow changes. Additional metadata elements were required to identify and link the three entities represented in version 3.0. Version 3.0 specifies a 1:1 relationship between the publication and data package module and a 1:N relationship to the data file module. This was an effort to bring back the publication as a first-class object within Dryad. It is important to note that version 3.0 was an idealized version of the Dryad application profile, and was never fully implemented due to constraints on the Dryad system architecture. In addition, it was determined to have few practical benefits to Dryad's users.

When Dryad was initially developed, there was no concept of the data package; instead, the domain model only included publications and associated data files. As Dryad grew, the idea of a data "package" was introduced. The records in Dryad that were formerly used to represent publications were changed to be data packages, though they still contain some information related to the publication. By recombining the publication and data package modules, version 3.1 represents a more feasible, scaled-down version 3.0, while still retaining the 1:N relationship between the package and file modules. With only two domain model entities in this version, fewer identifiers and relational elements were required to describe the contents of and relationships among the entities. Version 3.1 also demonstrated a consolidation of metadata elements related to the associated journal publication into a single citation metadata element.

As noted in Figure 1, version 3.2 of the application profile preserves the domain model of version 3.1, but includes changes in the metadata elements it represents. V3.2 includes elements for the manuscript number of the associated publication and a publication blackout release date, which corresponds to the date the associated publication will be released online.

4.3 Dryad Application Profile Version 3.2

The updated Dryad application profile is presented in Appendix A of this article and also published in Dryad (Krause et al., 2015). An example of three metadata elements is presented in Table 2. This table documents the namespace and name of the element as it is represented internally by DSpace; the element as it is represented externally as metadata is harvested by an API, a URI, a definition; the module in which the element is included, the obligation, and cardinality. Elements may be located in the data package module the data file module or both modules. The data package module contains 24 metadata elements from the Dublin Core, Darwin Core, and PRISM schemas, as well as from the Dryad namespace. Many of these elements, such as spatial coverage, subject, and scientific name, can be automatically propagated to the data file module. This reduces the effort required for the submitter to provide richer metadata at the individual file level. While the most common Dryad workflow is archiving data as part of the publication process, the repository is now supporting inclusion of data in the peer review process for several journals. This new workflow has had an impact on the set of metadata elements implemented by Dryad. For example, the metadata element `dcterms:manuscriptNumber` links a manuscript to its associated data package, allowing publishers to consider the associated datasets that underlie submitted manuscripts before they are published. The updated data file module

contains 21 metadata elements from the Dublin Core and Darwin Core schemas and the Dryad namespace. Data files are linked to the data package module through the `dcterms:ispartof` and `dcterms:relationhaspart` metadata elements, which point to the digital object identifier (DOI) of the linked modules.

TABLE 2: Selected Dryad Metadata Application Profile Elements, Version 3.2.

Internal Element Representation (DSpace):				dcterms:contributor.author	
External Element Representation (Metadata Harvesting APIs):				dcterms:creator	
URI:		http://purl.org/dc/terms/creator			
Definition:		Authors on publication / Authors of data submission			
Module(s):	Package & File	Obligation:	Required	Cardinality:	Repeatable
Internal Element Representation (DSpace):				dcterms:coverage.spatial	
External Element Representation (Metadata Harvesting APIs):				dcterms:spatial	
URI:		http://purl.org/dc/terms/spatial			
Definition:		Spatial description of the data specified by a geographic description and/or geographic coordinates			
Module(s):	Package & File	Obligation:	Optional	Cardinality:	Repeatable
Internal Element Representation (DSpace):				dcterms:coverage.temporal	
External Element Representation (Metadata Harvesting APIs):				dcterms:temporal	
URI:		http://purl.org/dc/terms/temporal			
Definition:		Temporal description of the data, as geologic timespan or dates of data collection/research			
Module(s):	Package & File	Obligation:	Optional	Cardinality:	Repeatable

5. Conclusion

This paper reports on efforts to align Dryad's application profile with current practice, and will be published as Version 3.2. Application profiles promote data sharing, interoperability, and linked data, which are all central to the overarching mission of Dryad. We performed a crosswalk analysis and diagrammed domain models to document and compare changes in the application profile. Over time, Dryad has changed the way it conceptualizes the relationships between data files, data packages, and publications. Furthermore, previous work on updating the application profile has revealed limitations in DSpace. Finally, examining which metadata elements and refinements have been added or deleted gives insight to which fields are the most crucial for archiving, preserving, and re-using data.

The data collected in this work is essential in outlining new goals for Dryad's metadata schema. Dryad's community has substantially expanded since its inception in 2007. In addition, the landscape of data repositories and archives has grown a great deal over past decade. New requirements for researchers regarding data deposition should be taken into consideration when deciding what information is collected from researchers about their data. The data collected through this effort will help inform future directions for metadata best practices across scientific data repositories.

As a next step, one of our goals is to publicly declare the Dryad-specific subproperties using the Dryad PURL domain. As indicated above, this paper reports on Dryad's work in day-to-day operational systems, but we have a long term goal to be more fully compliant with the DCMI and align with the Singapore Framework and the DCAM. This much longer-term goal will allow us map our labels onto RDF properties in order to achieve RDF Linked Data interoperability. In addition, we will perform a content analysis and examine a selected set of metadata schemas and

elements, such as DDI or PREMIS. In order to re-evaluate Dryad's functional requirements, it will be necessary to identify and consider new stakeholders (including journals, societies, researchers as both data depositors and data users, funders, and educators) and more complicated curation workflows. In order to determine users' needs, a next step could be to survey different types of users and follow up with more qualitative interviews. In addition, we will need to consider the increasingly diverse data formats and types that are used in the scientific domains represented in Dryad. New metadata elements may be needed to properly describe and preserve clinical data, social science data, and any other scientific data that Dryad could accept in the future. Finally, we will develop concrete objectives for implementing Dryad's metadata best practices, based on a deeper understanding of user needs and limitations of the repository.

Acknowledgements

We would like to acknowledge Ryan Scherle, Dryad Data Architect, and Thomas Baker, DCMI.

References

- Bibliographic Ontology. (2009). Bibliographic Ontology Specification. Retrieved from <http://bibliontology.com/>
- Carrier, Sarah. (2008). The Dryad Repository Application Profile: Process, Development, and Refinement (Master's Paper). University of North Carolina at Chapel Hill. Retrieved from <https://cdr.lib.unc.edu/indexablecontent/uuid:727a2712-f8f0-40a3-8876-e8e16b6d086d>.
- Coyle, Karen, and Thomas Baker. (2009). Guidelines for Dublin Core Application Profiles. Retrieved from <http://dublincore.org/documents/2009/05/18/profile-guidelines/>
- Creative Commons. (2015). CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Retrieved from <http://creativecommons.org/publicdomain/zero/1.0/>
- Darwin Core Task Group. (2015). Darwin Core. Retrieved from <http://rs.tdwg.org/dwc/>
- DCMI. (2007). DCMI Abstract Model. Retrieved from <http://dublincore.org/documents/abstract-model/>
- DCMI. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from <http://dublincore.org/documents/singapore-framework/>
- DCMI. (2012). DCMI Metadata Terms. Retrieved from <http://dublincore.org/documents/dcmi-terms>
- DDI. (2009). Data Documentation Initiative. Retrieved from <http://www.ddialliance.org/>
- Dryad. (2013). Metadata Profile. Dryad Wiki. Retrieved from http://wiki.datadryad.org/Metadata_Profile.
- Dryad. (2015). The Organization: Overview. Retrieved March 30, 2015, from <http://datadryad.org/pages/organization>.
- Dryad Data Repository. (2015). Retrieved from <http://datadryad.org/DSpace>.
- DSpace. (2015). DSpace. Retrieved from <http://www.dspace.org/>
- Dube, Jed. (2007). A Metadata Application Profile for the DRIADE Project.
- Dube, Jed, Sarah Carrier, and Jane Greenberg. (2007). DRIADE: A data repository for evolutionary biology. Proceedings of the 2007 Conference on Digital Libraries, Vancouver, BC, Canada, ACM Press, pp. 481. doi:10.1145/1255175.1255280
- Greenberg, Jane, Sarah Carrier, and Jed Dube. (2007). The DRIADE Project: Phased Application Profile Development in Support of Open Science. International Conference on Dublin Core and Metadata Applications, pp. 35–42.
- Greenberg, Jane, Hollie White, Sarah Carrier, and Ryan Scherle. (2009). A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata, 9(3), 194-212. doi:10.1080/19386380903405090
- Heery, Rachel, and Manjula Patel. (2000). Application profiles: Mixing and matching metadata schemas. Ariadne, 25. Retrieved from <http://www.ariadne.ac.uk/issue25/app-profiles/>
- Idealliance. (2015). PRISM Metadata Initiative. Retrieved from <http://www.idealliance.org/specifications/prism-metadata-initiative>
- Krause, Edward M., Erin Clary, Adrian Ogletree, Jane Greenberg. (2015). Data from: Evolution of an application profile: advancing metadata best practices through the Dryad data repository. Dryad Digital Repository. doi:10.5061/dryad.f0n35
- Library of Congress (LoC). (2015). PREMIS Data Dictionary for Preservation Metadata. Retrieved from <http://www.loc.gov/standards/premis/>
- Mannheimer, Sara, Ayoung Yoon, Jane Greenberg, Elena Feinstein, and Ryan Scherle. (2014). A Balancing Act: The Ideal and the Realistic in Developing Dryad's Preservation Policy. First Monday, 19(8). doi:10.5210/fm.v19i8.5415

- National Information Standards Organization (NISO). (2004). Understanding Metadata. Bethesda, MD: NISO Press. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- NCBI. (2012). Journal Publishing Tag Set. Retrieved from <http://dtd.nlm.nih.gov/publishing/>
- Nevile, Liddy, and Eva Méndez. (2015). Do We Need Application Profiles? Reflections and suggestions from work in DCMI and ISO/IEC. In DC-2015: Proceedings of the International Conference on Dublin Core and Metadata Applications. São Paulo, Brazil, September 1-5, 2015.
- Nilsson, Mikael, Thomas Baker, and Pete Johnston. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from <http://dublincore.org/documents/singapore-framework/>.
- Nilsson, Mikael, Alistair J. Miles, Pete Johnston, and Fredrik Enoksson. (2009). Formalizing Dublin Core Application Profiles – Description Set Profiles and Graph Constraints. In Sicilia, Miguel-Angel & Miltiadis D. Lytras (Eds.), Metadata and Semantics, pp. 101–111. Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-0-387-77745-0_10.
- Powell, Andy, Mikael Nilsson, Ambjörn Naeve, Pete Johnston, and Thomas Baker. (2007). DCMI Abstract Model. Retrieved from <http://dublincore.org/documents/abstract-model/>.
- White, Hollie. (2008). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. In International Conference on Dublin Core and Metadata Applications 2008, September 22-26, 2008, Berlin, Germany, pp. 202. Retrieved from <http://dcpapers.dublincore.org/pubs/article/view/936/932>.

Appendix A: Dryad Metadata Application Profile, Version 3.2

Internal Element Representation (DSpace):				dcterms:contributor.author	
External Element Representation (Metadata Harvesting APIs):				dcterms:creator	
URI:		http://purl.org/dc/terms/creator			
Definition:		Authors on publication / Authors of data submission			
Module(s):		Package & File	Obligation:	Required	Cardinality: Repeatable
Internal Element Representation (DSpace):				dcterms:coverage.spatial	
External Element Representation (Metadata Harvesting APIs):				dcterms:spatial	
URI:		http://purl.org/dc/terms/spatial			
Definition:		Spatial description of the data specified by a geographic description and/or geographic coordinates			
Module(s):		Package & File	Obligation:	Optional	Cardinality: Repeatable
Internal Element Representation (DSpace):				dcterms:coverage.temporal	
External Element Representation (Metadata Harvesting APIs):				dcterms:temporal	
URI:		http://purl.org/dc/terms/temporal			
Definition:		Temporal description of the data, as geologic timespan or dates of data collection/research			
Module(s):		Package & File	Obligation:	Optional	Cardinality: Repeatable
Internal Element Representation (DSpace):				dcterms:date.accessioned	
External Element Representation (Metadata Harvesting APIs):				dcterms:dateSubmitted	
URI:		http://purl.org/dc/terms/dateSubmitted			
Definition:		Date DSpace takes possession of item after a curator archives the item			
Module(s):		Package & File	Obligation:	Required	Cardinality: Non-Repeatable
Internal Element Representation (DSpace):				dcterms:date.available	
External Element Representation (Metadata Harvesting APIs):				dcterms:available	
URI:		http://purl.org/dc/terms/available			
Definition:		Date and time the package becomes available to the public on DSpace			
Module(s):		Package & File	Obligation:	Required	Cardinality: Non-Repeatable
Internal Element Representation (DSpace):				dcterms:date.blackoutUntil	
External Element Representation (Metadata Harvesting APIs):				N/A; Internal element only	
URI:		URI not assigned			
Definition:		A date after which the dataset will automatically archive itself (move out of publication blackout)			
Module(s):		Package	Obligation:	Optional	Cardinality: Non-Repeatable
Internal Element Representation (DSpace):				dcterms:date.embargoedUntil	
External Element Representation (Metadata Harvesting APIs):				N/A; Internal element only	
URI:		URI not assigned			
Definition:		Embargo date - a date after which the dataset will be made public			
Module(s):		File	Obligation:	Optional	Cardinality: Non-Repeatable
Internal Element Representation (DSpace):				dcterms:date.issued	

External Element Representation (Metadata Harvesting APIs):					dcterms:issued
URI:	http://purl.org/dc/terms/issued				
Definition:	Date of journal article publication				
Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dcterms:description
External Element Representation (Metadata Harvesting APIs):					dcterms:description
URI:	http://purl.org/dc/terms/description				
Definition:	Description of entity; In the data package module, refers to abstract of associated scholarly publication				
Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dcterms:description.provenance
External Element Representation (Metadata Harvesting APIs):					dcterms:provenance
URI:	http://purl.org/dc/terms/provenance				
Definition:	Information related to the origin and integrity of the file; history of custody of the item since its creation, including any changes successive custodians made to the item				
Module(s):	Package & File	Obligation:	Required	Cardinality:	Repeatable
Internal Element Representation (DSpace):					dcterms:format.extent
External Element Representation (Metadata Harvesting APIs):					dcterms:extent
URI:	http://purl.org/dc/terms/extent				
Definition:	Size of the file (bytes)				
Module(s):	File	Obligation:	Required	Cardinality:	Repeatable
Internal Element Representation (DSpace):					dcterms:identifier
External Element Representation (Metadata Harvesting APIs):					dcterms:identifier
URI:	http://purl.org/dc/terms/identifier				
Definition:	DOI of the Dryad entity (data package or data file)				
Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dcterms:identifier.citation
External Element Representation (Metadata Harvesting APIs):					dcterms:bibliographicCitation
URI:	http://purl.org/dc/terms/bibliographicCitation				
Definition:	Standard bibliographic citation of the associated scholarly publication				
Module(s):	Package	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dcterms:identifier.manuscriptNumber
External Element Representation (Metadata Harvesting APIs):					N/A; Internal element only
URI:	URI not assigned				
Definition:	Manuscript number of associated scholarly publication				
Module(s):	Package	Obligation:	Optional	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dcterms:identifier.uri
External Element Representation (Metadata Harvesting APIs):					dcterms:identifier
URI:	http://purl.org/dc/terms/identifier				
Definition:	URL which links to the web location of the Dryad entity				

Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:relation.haspart	
External Element Representation (Metadata Harvesting APIs):				dcterms:hasPart	
URI:		http://purl.org/dc/terms/hasPart			
Definition:		Record identifier for associated Dryad data file (doi:###/1 ; doi:###/2 ; etc.)			
Module(s):	Package	Obligation:	Required	Cardinality:	Repeatable
Internal Element Representation (DSpace):				dcterms:relation.ispartof	
External Element Representation (Metadata Harvesting APIs):				dcterms:isPartOf	
URI:		http://purl.org/dc/terms/isPartOf			
Definition:		Associated Dryad Data Package Identifier (doi:###) - the "root" doi of the package			
Module(s):	File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:relation.ispartofseries	
External Element Representation (Metadata Harvesting APIs):				N/A; Internal element only	
URI:		URI not assigned			
Definition:		Series name and number within that series, if available			
Module(s):	Package	Obligation:	Optional	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:rights.uri	
External Element Representation (Metadata Harvesting APIs):				dcterms:rights	
URI:		http://purl.org/dc/terms/rights			
Definition:		Statement regarding the rights held over the resource, e.g. CC0 (Creative, 2015)			
Module(s):	File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:subject	
External Element Representation (Metadata Harvesting APIs):				dcterms:subject	
URI:		http://purl.org/dc/terms/subject			
Definition:		Keywords associated with the Dryad entity			
Module(s):	Package & File	Obligation:	Optional	Cardinality:	Repeatable
Internal Element Representation (DSpace):				dcterms:title	
External Element Representation (Metadata Harvesting APIs):				dcterms:title	
URI:		http://purl.org/dc/terms/title			
Definition:		Title of entity (article, dataset, package, file, etc.)			
Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:type	
External Element Representation (Metadata Harvesting APIs):				dcterms:type	
URI:		http://purl.org/dc/terms/type			
Definition:		Entity type: article (package) or dataset (file)			
Module(s):	Package & File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):				dcterms:type.embargo	
External Element Representation (Metadata Harvesting APIs):				N/A; Internal element only	

URI:	URI not assigned				
Definition:	Length of Embargo (none, oneyear, custom)				
Module(s):	File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dryad:downloads
External Element Representation (Metadata Harvesting APIs):					N/A; Internal element only
URI:	URI not assigned				
Definition:	Number of times the data file has been downloaded				
Module(s):	File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dryad.externalIdentifier
External Element Representation (Metadata Harvesting APIs):					dcterms:identifier
URI:	http://purl.org/dc/terms/identifier				
Definition:	Unique identifier for related data in Dryad partner repository				
Module(s):	Package	Obligation:	Optional	Cardinality:	Repeatable
Internal Element Representation (DSpace):					dryad:pageviews
External Element Representation (Metadata Harvesting APIs):					N/A; Internal element only
URI:	URI not assigned				
Definition:	Number of times the webpage of a data file has been viewed				
Module(s):	File	Obligation:	Required	Cardinality:	Non-Repeatable
Internal Element Representation (DSpace):					dwc:ScientificName
External Element Representation (Metadata Harvesting APIs):					dwc:scientificName
URI:	http://rs.tdwg.org/dwc/terms/scientificName				
Definition:	Full name of the lowest level taxon to which the organism has been identified in the most recent accepted determination, specified as precisely as possible (may also specify other levels of biological taxonomy)				
Module(s):	Package & File	Obligation:	Optional	Cardinality:	Repeatable
Internal Element Representation (DSpace):					prism:publicationName
External Element Representation (Metadata Harvesting APIs):					prism:publicationName
URI:	http://www.prismstandard.org/specifications/3.0/PRISM_Basic_Metadata_3.0.htm#_Toc336960554				
Definition:	Name of publication associated with an item (i.e. journal name)				
Module(s):	Package	Obligation:	Required	Cardinality:	Non-Repeatable

Do We Need Application Profiles? Reflections and Suggestions from Work in DCMI and ISO/IEC

Liddy Nevile
(Retired)
Australia
liddy@sunriseresearch.org

Eva Méndez
LIS Departement / iSchool
Universidad Carlos III de
Madrid, Spain
emendez@bib.uc3m.es

Abstract

In this paper, the authors question the role and naming of ‘application profiles’ (APs). It is not a research paper but aims to foster a discussion that the authors think is pertinent. Both have been involved in the development and use of application profiles for some considerable time. This paper does not provide answers but aims to raise issues for others’ consideration. Essentially, the issues show that communities can share work easily through the interchange of APs but suggests that greater precision in their naming would be useful, and they may not always be necessary given the current state of RDF technologies.

Keywords: application profiles; metadata; metadata schemas; APs; MAP; RDF; discussion

1. Introduction

When someone you really respect makes a comment, even casually, it can fester for days. How about, “Why do you want an application profile? They are not necessary...” This comment was made in the context of developing a standard for Sub-Committee 36 of the Joint Technical Committee 1 of the ISO/IEC, a committee working on standards for ‘IT for Learning, Education and Training’ (ITLET). The target standard that had already been adopted and even mandated in Europe (ISO/IEC N24751) concerns accessibility of resources but was being significantly revised. The context included the development of a new comprehensive Metadata for Learning Resources (MLR) standard for ITLET (ISO/IEC N19788). The latter standard is very strictly Resource Description Framework (RDF) compliant and it is for education, so it also offers support for Learning Object Metadata (LOM) users, and for many in other DCMI related communities.

A similar comment was made by a student at the end of a course on “Metadata and Vocabularies”, after reading all the course material, recommended bibliography and so forth. He asked, “It is so common as it seems, the creation of so many application profiles? It seems that every single project of digital information service requires its own “customized” metadata schema? It has to be like that? It is not against the standardization that you said surrounds the metadata? On top, currently there are several schemas applicable to different projects so, I am wondering if it would not be enough choosing one of those standardized schemas.”

So, in different contexts, both authors have heard the same *‘Why do you want an ‘application profile’?’* or *‘Do you really need yet another ‘application profile’?’* The comment seems worth consideration in the context of another Dublin Core Metadata Initiative (DCMI) conference and set of tutorials, including one on application profiles (APs). In addition, the profiling process has expanded, given the proliferation and availability of RDF and Semantic Web technologies.

As Murtha Baca from Getty said, metadata standards are sometimes like toothbrushes, everybody thinks that they are a very good idea, but everybody prefers to use their own (Méndez, 2007). In this paper, the authors consider the role and naming of sets of terms for description of entities, of ‘application profiles’.

2. Metadata Application Profiles (APs)

A Metadata Application Profile, Metadata AP, or just MAP or AP can be understood from a number of definitions ranging from the more general one in Wikipedia to the more specific ones in the DCMI context. Wikipedia¹ defines an AP in the domain of ‘computer science’:

an application profile consists of a set of metadata elements, policies, and guidelines defined for a particular application. The elements may come from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata from several element sets - including locally defined sets. For example, a given application might choose a subset of the Dublin Core that meets its needs, or may include elements from the Dublin Core, another element set, and several locally defined elements, all combined in a single schema. An application profile is not complete without documentation that defines the policies and best practices appropriate to the application (Wikipedia, 2015).

At the time of writing, the Guidelines for the Dublin Core AP say:

A DCAP includes guidance for metadata creators and clear specifications for metadata developers. By articulating what is intended and can be expected from data, application profiles promote the sharing and linking of data within and between communities. The resulting metadata will integrate with a semantic web of linked data. To achieve this it is recommended that application profiles be developed by a team with specialized knowledge of the resources that need to be described, the metadata to be used in the description of those resources, as well as an understanding of the Semantic Web and the linked data environment (Coyle and Baker, 2009).

Thomas Baker et al. (2001) noted: “It is rare that requirements of a particular project or site can all be met by any one standard ‘straight from the box’”. Different metadata implementations may have different perspectives. Different information contexts, different content or different user requirements can motivate the creation of a Metadata Application Profile for local purposes. MAPs are the performance of the “*Think global, act local*” principle applied to metadata in domain-oriented digital information services. In fact, an early principle that drove the development of the DC Terms was that communities were likely to have domain relevant needs that may have little value beyond their context. On the other hand, they were surely interested in sharing their descriptions and therefore their term sets (APs) because that would assist with interoperability. This driving principle was embodied in the slogan ‘global interoperability and local specificity’ and cited many times in the early days of DCMI. In fact, APs were nurtured in a context where it was well-known that not every metadata system would be the same.

It should be pointed out that the work of Hunter and Lagoze (2001) led to the OAI (Open Archives Initiative) developments that partially solved the problem of sharing relevant ‘global’ metadata while the APs were conceived to solve the localisation problem.

In practice, it seems that nowadays communities develop APs for their domain of activity but it is safely assumed that when these are implemented locally, system developers choose what is of use from those community APs and locally they will, for sure, add some terms for local use (such as collection acquisition dates, or benefactors, for example). This practical approach to the use of APs has been described as a process in which the APs are used as “metadata building blocks” (Zeng and Qin, 2008). An AP may also be based on single schema but tailored to different user communities (Chan and Zeng, 2006); examples include the DC Library Application Profile (DC-Lib) used by libraries and library-related projects and applications or LOM-ES that explains the use of the Learning Object Metadata elements by the Spanish speaking community.

¹ The authors are quoting Wikipedia deliberately because that is where most people find meaning for such expressions. They are well aware of the many detailed, carefully defined definitions of application profiles that have been developed by authoritative entities, communities, in academic papers, etc.

Another way of dealing with this issue has been shown by the DCMI education and accessibility communities. Curiously, in concurrent meetings of the two groups some years ago at a DC Metadata Conference, the notion of metadata 'modules' was raised by those communities. They recognised that they had specialised needs but that what they wanted was just a small set of additional terms that could be used alongside the more general set of terms, or added to an existing AP.

So we start our conversation with local resource profiles, community resource profiles and modules for resource profiles in mind. Let us now analyze two particular AP contexts: DCMI, and ISO/IEC JTC1 ITLET.

3. DCMI context

3.1. Metadata for Education in the DCMI context

The aim of the first DCMI application profile, designed for education back in 1999, was to find a way the small DC element set could satisfy the needs of a specific community. Until then, DCMI work had been focused on developing a general set of elements for everyone to share. The value of the application profile, a slightly extended set of elements, was that it increased the value to a particular community by putting the focus on the properties of interest, following the DCMI idea of promoting 'global interoperability and local specificity'. This work was undertaken in the development of the Victorian Education Channel in Australia (Nevile, 2008, p 126).

The element set extension was done with the assistance of the then Director of the DCMI who considered three factors important. Any new term should:

- not redefine terms,
- not duplicate terms, and
- follow the dumb-down rule. (Nevile, 2008, p. 127)

Significantly, the new terms were to further describe the attributes of a given resource.

The exercise helped broaden the use of DC elements. The idea of application profiles was formalised in a paper written and published shortly afterwards by Rachel Heery and Manjula Patel (2000) where they specifically attached the concept of AP to data elements from different namespace schemas being combined by the implementor in a way that was optimized for a particular local application. Heery and Patel explained that application profiles are useful as they allow the implementer to declare how they are using standard schemas (APs). Thus there was recognition of a community developed schema (AP) and a local AP, often built from a combination of components of other APs. Again, the main aim is clearly to maximise global interoperability and, at the same time, local specificity. It was and is still also to enable better descriptions of an entity for the process of matching user requirements to available resources (or services).

Unfortunately, it seems in hindsight, the name 'application profile' stuck, without clarity about why. A number of different term sets were given the same name. For the purposes of this paper, the authors have distinguished between sets of terms for describing resources based on:

- agreement among a wide community for publication, and
- relevance to a particular context.

These sets can then be further defined as being determined to cover.

- fixed attributes of particular available resources but also, incidentally,
- user specified attributes of resources commonly thought of as search criteria in the discovery process.

The second distinction helps clarify that resource descriptions are always potential search criteria, or what has been described in other contexts as user needs and preferences (Nevile, 2005a, 2005b). Put simply, a resource provider might describe the date of publication in a

standard way and so a user can search for a resource with that date of publication in a corresponding way. There is nothing new in this, but the focus for a long time seemed to be on resource description, usually agreed among resource providers or organisers, and the use of the profile as search criteria was simply covered by saying the main use of the metadata was discovery.

This distinction is also useful because the growth and ability of 'search engines' that do not depend on what is commonly understood as metadata, or rather the lack of visibility of such search criteria, has led many to believe that search engines don't use metadata. Hopefully this myth has been rapidly and forcefully debunked recently by the spectacular growth and adoption of the work of schema.org. These new sets of metadata terms for 'all-the-web' retrieval systems is, at some point a 'déjà vu' for the authors, since schema.org revives the dream of qualified and precise search in the Web through metadata, like Altavista tried in the 90s. It is not the case even now that search engines necessarily use metadata in the same way, or the same metadata, as more formal traditional systems, but at least there is now an open dialogue between the two discovery system providers.

At the same time as the use of APs was evolving, the DCMI was working on what emerged as its 'abstract model' (DCAM) (Powell et al, 2007). Later in DCMI's life, Nilsson tried to find agreement between the DC metadata and LOM metadata in the educational context. He found that very different models led to very different forms of metadata and they could not be matched, so lossless interoperability was not possible. In general, he showed how difficult it is to match metadata from different structural models and argued for metadata to be interoperable it must be developed at least using compatible models, and developed a structured model to explain this (Nilsson et al, 2008). Following this work, metadata interoperability is considered by levels of interoperation. This model is illustrated in the following figure.

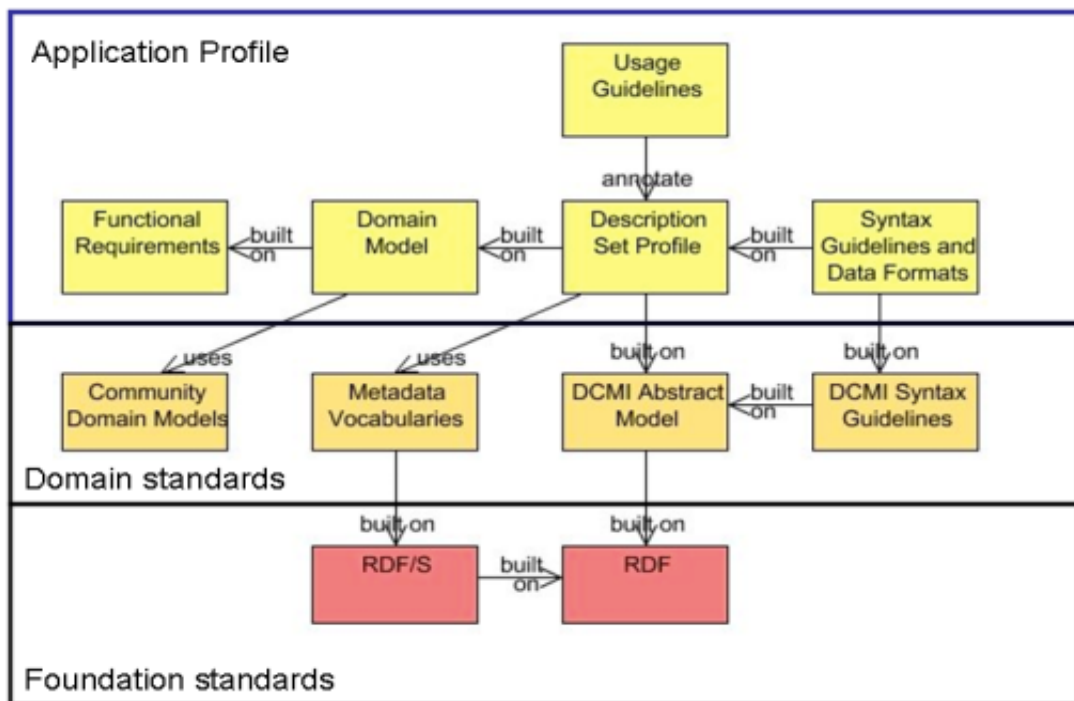


FIG. 1: the "Singapore Framework" for interoperability (Nilssen, et al, 2014)

At this time, the domain for description was always just considered as a 'resource'. (Perhaps significantly, DCMI did not distinguish between potential values that are now known as literals and non-literals.) But there was a rule, known as the one-to-one rule, which limited a description to a single resource. In the description or discovery of a resource, one might want also

to describe attributes of the person who made it, or the use to which it has been put. Curiously, the DCMI community overlooked that an image was a distinct entity, and for accessibility purposes may require a comprehensive description for those who could not see it. On the other hand, DCMI required a description of the resource and a second description of the person with the two descriptions linked by the term 'relation'.

Clearly, DCMI was a pioneering agency and the work was breaking new ground. The structures chosen were the best known at the time.

The integration of an image, or other object, into a resource was easy to live with when resources were published in what we might be described as a single entity form, despite really being a compilation (even including redundant parts like a long description of a diagram, tagged in the HTML as a longdesc). But today many resources are compiled 'on the fly' according to a wide set of requirements based not only on subject matter but location of the user, a number of previously exhibited behaviours, and more.

3.2. Metadata for Accessibility in the DCMI context

A short time after the forming of a DC Education Community, a DC Accessibility Community was formed. In the case of accessibility, a potential user needs to know such things as if there is a text alternative for an image, if a service can be controlled using only a keyboard or an assistive technology driven by a keyboard, or speech, perhaps. This means that a resource might need to take redundant forms, may not have all components assembled in a fixed format, and might need to be accompanied by an associated but not incorporated description of itself.

For more than a decade, the use of metadata to help solve the problem of inaccessibility of resources for people with disabilities has been pursued. There are guidelines for making resources that are, supposedly, accessible to everyone following what is called 'universal design principles' (W3C/WAI Web Content Accessibility Guidelines known as WCAG). Unfortunately, these guidelines are rarely followed successfully and even if they are, they do not satisfy everyone's needs simultaneously (e.g. Petrie & Bevan, 2009).

For some time, a term proposed for describing the accessibility of a resource was deemed unacceptable for technical reasons. Finally, a single term was adopted but the original hope that accessibility would become an important part of a DC metadata statement was not supported. This led to the work being taken to the ISO/IEC JTC1 context.

4. ISO/IEC context

4.1. Metadata for Education in the ISO/IEC ITLET context

ISO/IEC N19788 is the Metadata for Learning Resources (MLR) standard. The MLR is very detailed in its ways of defining application profiles (Part 1) and includes several APs. It has many parts and bridges earlier practices in both ISO/IEC's provision of what we now call metadata terms, and the practices associated particularly with older database systems and the hierarchical structures of the LOM.

The interoperability of the MLR comes not just from working carefully with the earlier practices of describing electronic resources using a sort of document object model, but the fact that today not only resources as they have been traditionally known are to be described. There are people associated with the development and publication of resources; there are services and online communities. All of these things are connected in a web of digital descriptions so users can have very different points of contact with that web and it refers to objects that are digital but also physical or merely conceptual.

N19788 is not necessarily easy to read in its full form, but that is not necessary for its use. It is very helpful in that it does provide full explanations of its techniques. There has been considerable effort put into diagrammatic representations, examples in pseudo code, and bindings

that can be simply adopted and used. The interoperability of the MLR depends, in fact, on complexity that is buried in what appears as an elegant framework. 'Under the covers' techniques have been developed to ensure that terms are easily accessible online, that internationalisation is fundamental, and more.

The MLR offers global interoperability by specifying how to do a number of things but strongly supports local specificity in terms of extensions, options, etc. The MLR's application profiles provide an initial set of core terms that can be used to describe educational resources, and these are effectively the 15 DC simple terms (limited currently by their domain to 'learning resources'). This application profile is in Part 2 of the standard. Part 4 of the standard offers a few specific terms to describe technical aspects of a resource and there is another application profile in Part 5 that has been developed by a community of educators to describe what might be considered the pedagogical aspects of learning resources.

The MLR goes on to include sets of terms for descriptions of, for example, the role of a person associated with the development of the learning resource, or of a resource that is, in fact, a metadata description of a learning resource. Such a term does not aim to support description of the resource itself, but an attribute of the person who has been described in association with a resource (or more particularly, the role of a person who has been so described). In this case, we think of the chaining of descriptions to link the various types of descriptions to form a web of information about the resource, significantly using RDF and data linking techniques, but it can be done however the user chooses.

The MLR has attempted to bridge the gap that emerged as technologies have moved from standard databases to more and more fluid systems. A considerable amount of what is in the MLR is concerned with this. The result, however, is that terms can be used and combined in very flexible ways.

4.2. Metadata for Accessibility in the ISO/IEC JTC1 Context

Today, many in the accessibility community have adopted the additional approach of profiling the needs and preferences of users, especially those with disabilities. The aim is for so compilations of resources to be matched to an individual user's stated needs and preferences so the resources are 'perceivable, operable, usable and robust' for them (WCAG). This, of course, means simply that if the relevant properties can be identified, they can be used for resource description by resource providers (or others) and for resource discovery (in search requests or automatically by systems). A delivered resource may not be accessible to another individual in the same form, or even to another user with a similar disability. The required form of the resource is for an individual user to define.

The AccessForAll approach, as it is known, was first proposed by Jutta Treviranus. It is simply a name for ways to describe an individual user's functional requirements for a resource that can be matched when a resource is being delivered. To metadata communities, it is a very normal metadata activity but somehow has not been recognised as such by a number of those who want it, and so, even after about 7 years of work, they have not managed to agree on what could be described as a simple AP - possibly all that is required!

A characteristic of AccessForAll metadata, as proposed, is that the attributes or properties of a resource are described using a set of terms which is the same set for a resource and for the search for the resource except that, in the latter case, there is no clear identity of the resource being described - its identity is being sought. That is, the same terms can be used but the identity of the resource is not specified in the latter set. It is appropriately described as a *module* of metadata. It challenges the idea that an AP is a set of terms that describes a resource. The description of the needs and preferences of a person, expressed as metadata, does not describe the person. In fact, a single person may have a number of stored accessibility modules describing the functional requirements that they use at different times, in different locations, and even according to different purposes.

Giving the terms for a search a different name from the terms for a description should settle this easily. It should not be a show stopper. It means simply that the identifier should be optional, whereas it has always been mandatory for a DCAP. There is nothing special about a module that describes attributes related to accessibility: the same idea can be used for many types of customisation of resources useful to anyone. This is generalisable as 'inclusion', the preferred way of avoiding discrimination.

The values set for the term set is, itself, an entity and that, of course, can have an identity in the form of a URI, or otherwise. It can have lots of other attributes as well and they too can be described in a value set.

The AccessForAll metadata approach has received significant funding for many years and is often considered to be exemplified by the project known as GPII (Global Public Inclusive Infrastructure). Sadly, despite the funding and academic papers and other peripheral successes, the simple matter of providing an 'application profile' for accessibility has not yet been achieved.

The AccessForAll idea is to have a profile of the needs and preferences of an individual user (could be anyone but should, at least, be inclusive of any person with disabilities) and to match those needs and preferences, strictly described as functional needs, to resources. If a resource is well-developed and has available components, possibly redundant, that can be combined to make all its essential content available to the individual user, the useful combination should be delivered.

If components are not accessible, for example an image is not also described in text, an alternative resource might be located or created to serve this purpose. The useful component can be linked to the original using the metadata. The whole matching exercise is known to provide what is an 'accessibility service' by constructing an 'accessible resource' and this can be a dynamic process, with cumulative accessibility.

5. RDF and APs

Developing the MLR (Metadata for Learning Resources) has provided an opportunity for re-thinking the original metadata and application profile ideas in a modern context, specifically in the context of an RDF world. Gilles Gauthier has provided an image of a web of descriptions as it might be for a particular learning resource of interest (Figure 1).

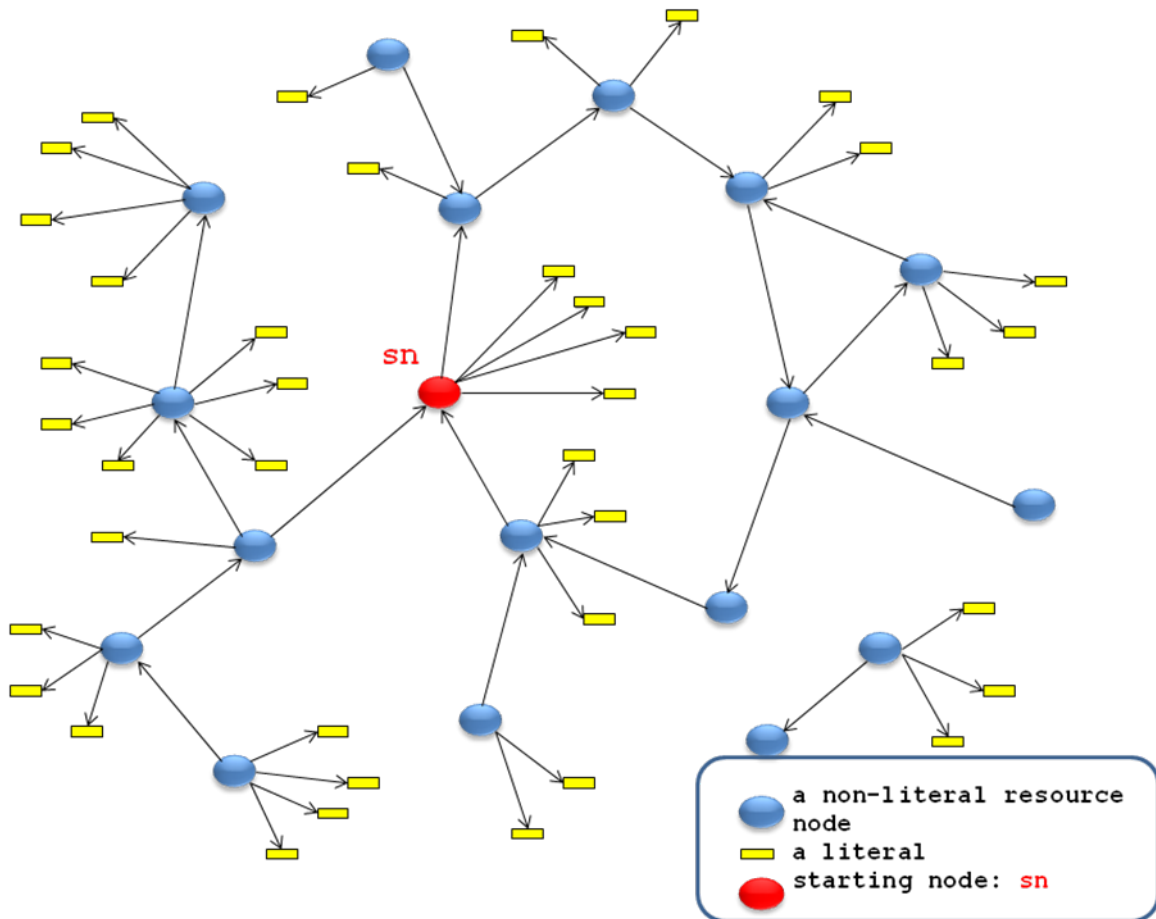


FIG. 2: an RDF graph showing an extensive web of RDF triples.

Using RDF triples, one could construct an impossibly huge web of descriptive triples for any resource, almost, but this is not always totally practical. If a resource description is to be useful, it may not matter from where the triples come, assuming they are reliable, and they can be all joined up but limited by a set of delimiters. These would be rules to say just how much information is wanted. The original map shown above is shown after a set of delimiters have formed the map that the particular user wants to work with (Figure 2).

So here is a question: Is what we see in Figure 2 an application profile? Is it possible for an application profile to be a set of delimiters? The MLR has lifted some of the restrictions earlier encountered in a way that is being done by many others. The current authors would like to suggest that the focus on ‘application profiles’ that has been useful in the past may benefit from some sort of re-thinking in the light of such new possibilities.

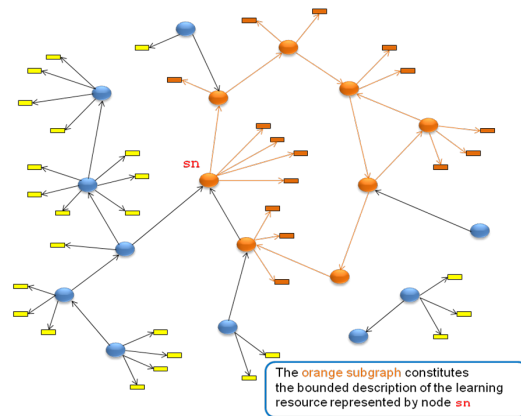


FIG. 3: the RDF graph from Figure 1 with a set of delimiters applied showing how it is thus ‘contained’.

6. Discussion

During the last decade, there has been extensive work on the practices for using APs, perhaps well-exemplified in the work of Curado and Baptista (2013) that provides a carefully researched process for developing an AP. This current paper, recognises this work, best practices for system developers, but also the work of communities that have resulted in APs such as the LOM (Learning Object Metadata), the LRMI (Learning Resource Metadata Initiative), the MLR in the educational domain and more in other information domains.

The authors are not sure why APs have a special name, which is cryptic for people who are not metadata-literate. Isn't an AP simply the set of terms that a user, whatever their role, chooses to use? Can't the set of terms be just that, and the ‘development of an AP’ be simply recognised as the ‘use of metadata terms’, a general activity? Couldn't the DCMI workshops on DCAP be known simply as workshops about how to use metadata?

Perhaps the value of the focus on ‘application profiles’ is that it helps people distinguish between the set of terms that can be used to describe a resource and the set of values for those terms. This difference is a significant problem for some people. For others, the idea of loose terms is a problem. They want to think of metadata terms as they thought of fields in a data base. They want them neat and controlled, probably in the same place, verified not just technically but by some authority. This is not the world in which metadata lives today. schema.org developers, some of the major search engine companies, have publicly stated that what they use will emerge according to what others use. Terms defined as part of schema.org that don't gain popularity will be ignored. schema.org can safely adopt the position of allowing terms to see if they work. The unused terms will not do harm. Terms published and not used will be just that.

The authors' interest in the name ‘application profile’ is perhaps also motivated by the fact that the expression is curious, even funny when translated into some other languages, and sometimes confusing.

The original specification of the domain of a DC term was very loose, simply ‘resource’. Given the different aspects of the resources we use today, there are a number of different parts of a resource that may need description so the domains will not be the same for all the descriptors in a useful metadata set. Already the capacity to handle this was proving a strength of Resource Description Format (now Resource Description Framework, RDF) metadata. That a resource may, in fact, be delivered in different forms or manifestations, according to user preferences, device types, etc., was not yet an issue.

The early RDF work anticipated that chaining of descriptions would be useful but, at the time, it was not well supported by software or implementations. RDF was not universally trusted in the late 90s - it seemed to be a folly for a small number of ‘academic’ types, semantic geeks and perhaps, as was often said, ‘people with comfortable shoes’. There was strong concern that it would go away so should be treated with caution. Time has shown something else. RDF is very

well established now and substantiates most of the Linked Open Data projects, but it is still relatively poorly understood. Many with a background in database work have a strong sense of structure and formalities that are not always compatible with RDF use. It is very hard for many to let go of those structures and leave organisation to the implementing systems. Nowhere has this been more problematic than in the ISO/IEC SC36 metadata work. There are representatives of technologically advanced nations who themselves do not endorse RDF and the Semantic Web. There perhaps even more who do not even know what is the difference between old databases and the Semantic Web... But today there are billions of RDF triples in use; hence the question being asked in this paper. Do we still need 'application profiles'? Do we still need to use that name? Might we want to advise people just to 'use metadata' and even share it, or develop it collaboratively? Alternatively, might we want to be more specific and recognise the various kinds of metadata profiles? What level of standardisation or at least community endorsement does an AP need? Most of the metadata models are developed as 'standards'.

Determining how to describe Japanese manga has offered a number of examples of resources that fall outside the norm. Manga, originally Japanese comic format (but often created with significant adult themes), is very different from standard literature. Like other comic series, characters re-appear in subsequent comics in a series. But more importantly, like literature, manga has a grammar. There are quite formal ways of signifying emotions and actions in manga (manga creators study for several years at the Japanese University of Manga in Kyoto). A useful way to think of manga is to compare it to ballet and other forms of dance. Metadata for the description of manga is complex - an elegant set of descriptive terms that includes the various attributes in the Functional Requirements for Bibliographic Records (FRBR) is the solution emerging from the work of Shigeo Sugimoto and his students in Tsukuba, Japan (Mihara, Nagamori, & Sugimoto, 2012). Such a complex AP would be beyond the average user to develop but once it has been established, it can be used easily. But how should such a set of terms be described? as an AP? What about calling it a 'manga profile set'? Wouldn't such a name be helpful?

Similarly, developing the MLR has been a very technically challenging exercise but the result is something that can be used by people with few computer development skills but good cataloguing skills, or maybe without them. The idea then is that the expertise to determine an appropriate set of terms and potential values for description of the wide range of resources may well be beyond the average user, but useful to them. But what is the MLR? it offers a number of APs but other terms as well. Metadata that mixes terms from the MLR and with others that conform to the MLR will be considered MLR metadata. Does it not make sense to talk about a set of educational metadata terms? In this case, given terms are defined both by text definitions (traditional term definition) and by constraints on RDF triples (newer term definition).

7. Conclusion

So, what is an application profile is not clear, according to the authors of this paper. It is a wide-ranging concept that is perhaps not even useful any more. Without reaching a conclusion, the authors hope to have stimulated some useful thinking and that some of the questions asked in this paper will lead to timely and useful discussion.

References

- Baker, Thomas et al. (2001) "What Terms Does Your Metadata Use? ApplicationProfiles as Machine-Understandable Narratives." *Proc. Int'l. Conf. on Dublin Core and Metadata Applications 2001*. Retrieved May 05, 2015 from <http://dcpapers.dublincore.org/pubs/article/view/654/650>
- Coyle, Karen, Thomas Baker (2009). *Guidelines for Dublin Core Application Profiles*. Retrieved May 05, 2015 from <http://dublincore.org/documents/profile-guidelines/>
- Curado Malta, Mariana and Ana Alice Baptista (2013). *A Method for the Development of Dublin Core Application Profiles (Me4DCAP v0.2): Detailed description*. *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013*. Retrieved May 05, 2015 from <http://dcpapers.dublincore.org/pubs/article/viewFile/3674/1897>

- Chang, Lois Mai, Marcia Lei Zeng (2006). *Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level*. D-Lib Magazine, 12 (6) Retrieved May 05, 2015 from <http://dlib.org/dlib/june06/chan/06chan.html>
- Heery, Rachel, Manjula Patel (2000). Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne*, September 2000 (25). Retrieved May 05, 2015 from <http://www.ariadne.ac.uk/issue25/app-profiles>
- Hunter, J. & Lagoze, C. (2001). Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles. WWW10. May 1-5 Hong Kong, ACM 1-58113-348-0/01/0005. pp. 457-566
- Méndez, Eva (2007). Singapur, Metadatos y cepillos de dientes. *Anuario Thinkapi*, pp. 63-66. Retrieved May 05, 2015 from <http://dialnet.unirioja.es/descarga/articulo/3190887.pdf>
- Nevile, L. (2005a). Adaptability And Accessibility: A New Framework in Proceedings of OZCHI 2005, Canberra, Australia. November 23 - 25, 2005. ACM (the Association for Computing Machinery) Digital Library. Retrieved December, 9, 2008 from <http://delivery.acm.org/10.1145/1110000/1108413/p21-nevile.pdf?key1=1108413&key2=2075188221&coll=ACM&dl=ACM&CFID=14456173&CFTOKEN=26161489>
- Nevile, L. (2005b). Anonymous Dublin Core Profiles for Accessible User Relationships with Resources and Services, in DC 2005 Conference. Retrieved May 05, 2015 from <http://dcpapers.dublincore.org/pubs/article/view/804/800>. Republished as Anonymous Dublin Core Profiles for Accessible User Relationships with Resources and Services in *New Technology of Library and Information Service*, 1/2006(132), pp. 17-24.
- Nevile, L. (2008). Metadata for user-centred, inclusive access to digital resources: Realising the theory of AccessforAll Accessibility. Retrieved May 05, 2015 from <http://researchbank.rmit.edu.au/eserv/rmit:8686/Nevile.pdf>
- Nilsson, M., Baker, T. & Johnston, P. (2008). Interoperability levels for Dublin Core metadata Retrieved July 10, 2008 from <http://dublincore.org/architecture/wiki/InteroperabilityLevels> Archived 2008-10-15 by WebCite® at <http://www.webcitation.org/5bZmdmXxZ>
- Nilsson, M., Baker, T. & Johnston, P. (2014). The Singapore Framework for Dublin Core Application Profiles. Retrieved May 05, 2015 from <http://dublincore.org/documents/2008/01/14/singapore-framework>
- Petrie, H. & Bevan, N. (2009). "The evaluation of accessibility, usability and user experience" in *The Universal Access Handbook*, C Stepanidis (ed), CRC Press, 2009.
- Powell, A., Nilsson, M., Naeve, A., Johnston, P. & Baker, T. (2007). DCMI Abstract Model. Retrieved October 18, 2007 from <http://dublincore.org/documents/2007/06/04/abstract-model/> Archived 2008-10-15 by WebCite® at <http://www.webcitation.org/5bZnACFp3>
- Mihara, T., Nagamori, M., & Sugimoto, S. (2012). A metadata-centric approach to a production and browsing platform of Manga. In *The Outreach of Digital Libraries: A Globalized Resource Network* (pp. 87-96). Springer Berlin Heidelberg.
- Wikipedia (2015), quoting "Dublin Core metadata glossary" archived (<http://web.archive.org/web/20060621074245/http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml>) from the original on 21 June 2006. Retrieved May 05, 2015 from http://en.wikipedia.org/wiki/Application_profile#cite_note-Dublin_Core_glossary:_Application_profile-1
- Zeng, M. & Qin, J. (2008) Metadata. ALA Neal-Schuman.

Web References

- ISO/IEC JTC1 SC36 ITLET http://www.iso.org/iso/iso_technical_committee?commid=45392
- MLR Part 1 Framework http://standards.iso.org/ittf/PubliclyAvailableStandards/c050772_ISO_IEC_19788-1_2011.zip
- RDF <http://www.w3.org/RDF/>
- schema.org <http://schema.org>
- WCAG <http://www.w3.org/TR/WCAG20/>

Language-acquisition inspired sustainability modelling for application profiles

Emma Tonkin
University of Bristol
United Kingdom
e.tonkin@bristol.ac.uk

Abstract

The ongoing accessibility of digital material is challenged by the constantly changing environment in which it exists. In particular, application profiles are threatened by a number of factors such as loss of context, social change and linguistic change. In this paper, we draw on observations taken from a number of application domains to build simple mathematical models for community growth and change, to explore the impact of community structure on the sustainability model required for application profiles over time. Finally, we discuss the use of similar models in evaluating application profile sustainability in general, and lessons to be drawn for DCMI.

Keywords: application profile; sustainability; user community; implementation

1. The application profile

The concept of the application profile is widely used in the world of Dublin Core, and expresses the idea that metadata, as it is experienced by its user communities, is situated within its context of use. To quote Heery and Patel (2000), 'implementors use standard metadata schemas in a pragmatic way'; those making day-to-day use of implemented systems are very likely to make use of the system to fulfil their task to the greatest extent possible. Ideals of semantic purity seldom survive exposure to the furnace of everyday pragmatism.

Application profiles reflect interdisciplinary boundaries and 'ways of seeing' (Berger, 1972) and may therefore be viewed as artefacts worthy of evaluation and exploration in their own right. Much as Olson (1998, 2001, 2002) makes use of library catalogues in the exploration of 'the cartography of marginalised domains' (Olson, 1998), so the creation and use of metadata application profiles provides a mirror through which practitioners may view institutional and individual practice.

Few of us explore the mirror images that application profile development makes available to us, with justification, given that these are functional artefacts intended to support the development of a computer-supported system that solves a problem. Invisibility could be said to be a design goal in application profile development: when the user finds themselves wondering about an application profile, it may plausibly imply that the profile has failed to achieve a stated goal. For practitioners, an application profile attracts little interest, beyond the question of whether it adequately reflects the needs of those working in the domain or with the system. Far less do practitioners find their gaze trapped, like a mythical Narcissus, in the reflection of their work. Indeed, it could be said that what Heery and Patel (2000) refer to as 'standards-makers' have a far greater propensity to the Narcissan fascination with reflection of self, being more often driven by the search for integrity, consistency and contemporary ideals of design and implementation.

1.1. Sustainability and the application profile

Application profiles represent a localisation of terms drawn from one or more relatively decontextualised concept spines (namespace schemas). Where parent resources may be viewed as subject to the pressures of social and cultural change (Kapitzke, 2001), the sustainability of the resource is called into question. Although metadata is one of the key pillars upon which data preservation efforts rest, it is the metadata that may cause greater concern than the preservation of data objects themselves; metadata is expensive to generate and its use can be expected to rely to some greater or lesser extent on the availability of standard components, such as metadata registries, or other components of the OAIS functional model (Day, 2002). Such components are reliant on a level of ongoing support and continuity, and (as shared resources in a broadly shared context) on a coherent multiorganisational or even multinational commitment to collaboration.

1.2 Evolution of an application profile

Application profiles themselves, representing a form of internationalisation or localisation, may be expected to suffer from the ongoing processes of change imposed by the drivers acting on that domain. Some result from changes within the organisation or community; some are the consequences of external change. Consider for example:

- external or internal political or strategic mandates
- staff turnover within an organisation
- organisational structure and project lifecycle
- changes in social attitudes

It may be gathered from this that the speed of change imposed on an application profile is not uniform. It is dependent on the characteristics of the community that the profile is designed to support. The maintenance requirements, and consequentially the sustainability of an application profile, can be expected to depend on situational and environmental factors. This broader set of contextual factors also includes the commercial, legal, regulatory and market context, which is referred to by Messerschmitt and Szyperiski (2003) as the 'software ecosystem' in which any given system can be seen to operate.

Given that this short section covers a large number of factors, we cannot hope to explore all of these issues within a single paper; hence, we narrow our focus to a specific question: what is the effect of rapid change in user community on the rate of change imposed upon, and hence the sustainability of, an application profile?

1.3 Semantic evolution, shift, drift and change

In this paper, the mutability of various aspects of the system is considered. In particular, we explore the factor of *semantic evolution*, informally definable as a change in some part of a system, which typically results in a shift in the way in which a term or concept is understood. These concepts originate in linguistics, where they are primarily used in the fields of sociolinguistics or historical linguistics to describe variation in the use of spoken or written language over time or distance.

In simple terms, a semantic change is a change in the way in which terminology is used; when we begin to use the word 'cool' to mean 'I agree' or 'excellent' rather than to describe a temperature beneath that of 'hot', then we have implemented a semantic change. Semantic evolution is understood to be a destabilising factor in software ontologies (Cudré-Mauroux et al, 2006). The term 'semantic drift' is sometimes used, as with Gulla et al (2010), who define the term as 'the gradual change of a concept's semantic value as understood by the relevant community'. Gulla et al divide the term into two main areas: *intrinsic* and *extrinsic* drift, in which an intrinsic drift reflects change with respect to other concepts within the same frame of reference (such as an ontology or similar structure), and an extrinsic drift represents change with respect to the real-world referent.

Semantic change can take various forms and have been modelled by a number of researchers (Bloomfield, 1933). To a certain extent, models mirror the well-known thesaural relations of broadening (increasing the breadth of use of a term) and narrowing (reduction in the breadth of use of a term), although many other dimensions of semantic change are tracked by various models.

Baruzzo et al (2009) remark that 'preservation of [digital] information is about maintaining the *semantic* meaning of both the digital object and its content'; social change plays a significant role in patterns of change observed within the user community, and hence user requirements evolve over time. For Baruzzo et al, semantic evolution occurs within three *evolution dimensions*, including

- the informational domain (metadata and knowledge organisation)
- the technological domain (technological infrastructure, human-computer interaction issues and information transfer issues)
- the social domain (human and organisational factors, legal, social and procedural change)

We may hypothesise that semantic change is particularly likely to occur in situations in which items or systems are not often accessed or used. As Kanhabua (2013) states, items that are not in active use may require a form of 'recontextualisation' in order to retrieve the item as it would originally have been perceived. That is, in plainer terms, if we cannot remember what something was supposed to mean or how it was intended to be used or perceived, we will have to spend time and effort developing and testing a hypothesis and resolving any issues encountered along the way. Change that remains unnoticed is more likely to be disruptive, since it is unremarked and consequently uncompensated.

2. Methods: modelling for sustainability

In order to understand the likely development path of a domain, it is common to make use of a simulation-based modelling approach. Due to the problematically high complexity of software systems, models are generally designed with the intention of a simplified representation of some subset of the domain. The interdisciplinary nature of sustainability evaluation means that models are often interdisciplinary in focus, reach and usage; There are a large number of modelling approaches designed or applied to support sustainability evaluation. For example, Penzenstadler et al (2012) reviewed available literature for sustainability in software engineering, identifying a number of models proposed by authors over time.

Models proposed include, amongst others:

- conceptual and reference models designed towards specific areas of sustainability, such as the GREENSOFT model (Naumann et al, 2011), which are themselves typically used as inspirations for specific modelling instances rather than serving as operative models in their own right; the GREENSOFT model, for example, powers various subprocedure models applied through creation and manipulation of UML sequence diagrams, guidelines, checklists and so forth;
- agent-based models (Axelrod & Tesfatsion, 2006);
- evolutionary theory (Safarzyńska et al, 2012);
- probabilistic approaches making use of Bayesian networks (Calero et al, 2012);
- ontology-based ecosystem modelling (Franch et al, 2013);
- goal-oriented techniques for stakeholder modelling, using modelling languages such as *i**, essentially a graph-based modelling approach (Cabot et al, 2009);
- cognitive modelling and fuzzy inference (Rajaram & Das, 2010).

Selecting an appropriate model clearly depends on the model's purpose: in the words of Box (1987), 'Essentially, all models are wrong, but some are useful'. Prior to choosing a model, we must therefore define our purpose, which, in our case, is the development of a model that models

the effect of factors identified in Section 1.2 of this paper on the evolution of the application profiles.

In this instance we explore the use of a model that to our knowledge has not previously been used for the purpose of sustainability modelling, but which has previously been used for the analogous purpose of computationally modelling the acquisition of language: a straightforward model of language acquisition. A discussion of computational modelling in language learning may be found in Kaplan et al (2008), although detailed evaluation of the model's original purpose exceeds the scope of this paper.

2.1 A simplified model of language acquisition

For the purposes of this paper, we apply a simple model based loosely on Niyogi (2006) and comparable to that discussed by Kaplan et al (2008). We make the following assertions: firstly, we accept that the linguistic knowledge and behaviour that underlies an application profile can be described as a formal system (Niyogi., p.37), and that human agents hold a range H of these systems. In order to successfully learn any given system h under this model, an individual must be exposed to events in which the term is used by a competent speaker of h . Secondly, we assert that h may be learned completely by an agent new to this system, by means of learning all terms used within the system. Finally, the process of learning a given term depends on two factors: exposure to at least one situation in which the term is correctly applied, which provides an opportunity to learn, and on the learnability l of the term. Learnability here refers to the probability that a given event in which an individual is exposed to a usage of the term will lead to a successful acquisition of the term. An individual who has been successfully exposed to all terms within h may be viewed as a competent user of h .

This model is unrealistic for several reasons: it discounts the possibility that a number of variants of any given system h may exist, whereas in practice variation within a formal system is likely to occur. Similarly, it presumes that a system must be completely learned in order for a user to be classified as competent. Additionally, it presumes that agents are entirely dependent on exposure to events in which terms are used to develop an understanding of how terms should be used. In practice, agents may also learn from documentation, although the learnability of examples given in documentation may diminish over time, as Kanhabua (2013) suggests, hence decreasing the accessibility of the material and reducing the efficacy of the documentation.

3. Qualitative case study: Continuous and discontinuous communities

In this section, we apply the model described in Section 2, above, to two sample cases. The first case describes a close-knit team with low staff turnover, which regularly makes use of an application profile. This case is similar to that found in many museum or archive contexts, in which continuity of practice is a significant factor. The second case describes a team which establishes an application profile, uses it for a certain period of time and then disbands; the data is then retrieved by another team, which attempts to make use of the application profile in question. This case resembles that often found in scientific research contexts, in which a project-driven team works for a certain period of time; the data and metadata created is preserved, and may well be retrieved at some later date for use in another context, such as a rapid innovation event or a later research project.

As a further simplification, we assume that the learnability of all terms in each case is total (i.e. $l=1$). Both case studies are dependent on the probability of the learner agent receiving evidence about terms in set h . Consequentially, this behaviour can be represented by a Markov chain.

We assume an application profile of ten terms, t_1 - t_{10} . We assume an equal probability that any of these terms are used, although in practice, evidence of the active usage of application profiles shows us that some terms are used markedly more frequently than others (Dushay & Hillmann, 2003). Hence, a learner with moderate competence is more likely to be confident on commonly used terms.

3.1 Application profile acquisition in a highly connected team context

A learner has a high probability at any time of encountering a learning event. We model the transition matrix accordingly; a section of the full transition matrix is shown below, showing the initial state and the final absorbing state, in which a learner has correctly grasped all terms and has therefore fully completed the learning process.

$$P(x) = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0.9 & 0.2 & \cdots & 0 \\ 0 & 0.8 & \cdots & 0 \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

In accordance with the high probability of observing events from which they can learn (i.e. expert uses of the terminology), the learner rapidly begin to learn terms. Once the process has begun, they learn rapidly.

3.2 Application profile acquisition in a sparsely- or disconnected context

In a context in which no learning events take place, it is clearly impossible for a new learner to become fluent, since no term learning events can occur. There is no need to model this explicitly since it is trivially clear that the transition matrix is empty, and cannot lead the learner to a productive state.

Instead, we model a context in which learning events take place with relatively low frequency (a 1:10 ratio relative to the first community). Whilst this still permits learning, it reduces the probability that any given simulation timestep will be *productive* (that the learner will learn something new during that timestep). We therefore alter the transition matrix to take account of this assumption.

$$P(x) = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0.99 & 0.02 & \cdots & 0 \\ 0 & 0.98 & \cdots & 0 \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

We expect this to reduce the learner's learning rate relative to the highly connected case.

4. Results and discussion

For each case in Section 3, we apply the transition matrix to a starting vector representing the initial state of our learner: $[1 \ 0 \ 0 \ \dots \ 0]$. The transition matrix is re-applied until equilibrium is reached, which in the case of this model concretely means that the learner has completely learned the terms in the application profile.

Graphically evaluating the results of cases 3.1 and 3.2 in figure 1, we find that the results comply with our expectations, showing that our learner picks up term usage rapidly in the connected state, and slowly in the sparse state. We have also observed that a learner without opportunity to learn will not acquire terms in this model, although in practice alternative learning strategies would undoubtedly be applied, such as learning from available documentation or available exemplars of use. Whilst an explicit model of this is beyond the scope of this paper, we remark that reduced learnability would have the result of slowing the process of learning further, stretching the S-shaped curve.

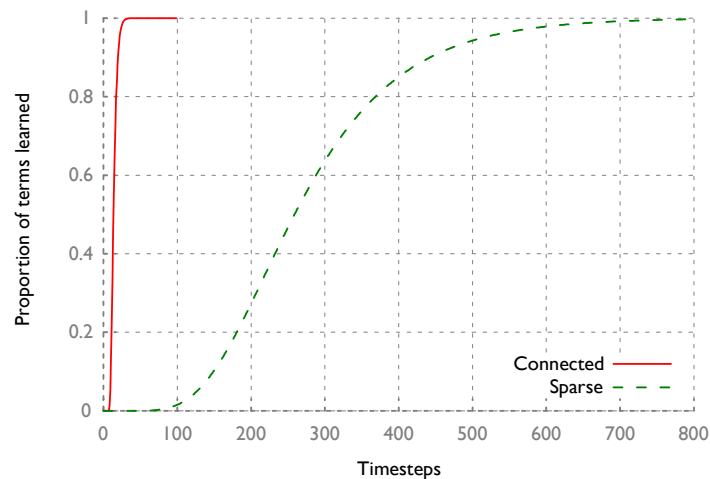


Figure 1: Model of term acquisition in connected and sparse community states

A further point of note is that, while the term acquisition rate varies significantly between the states, the curve itself does not. The S-shaped curve occurs in both states. Similar curves appear in many discussions of language change (see Niyogi 2006, pp. 29–30).

If insufficient exemplars or documentation were available, we would find an extremely low probability that a learner would successfully learn the usage of certain terms. This could have a number of possible effects. A learner might simply fail to learn any usage of the term, effectively truncating h by excluding the term entirely. Alternatively a learner may learn a differing interpretation of the term, resulting in the learner developing (and propagating) a variant form of the termset h , which we might refer to as h' . In the event that this occurs, this learner has experienced and will propagate a semantic change within the termset.

4.1 Discussion

We have shown that the availability of an active community has a significant effect upon the learner's ability to develop an understanding of terminology. We have also discussed that a would-be learner without an active community from which to learn must rely on available exemplars which act to demonstrate terminology in use, as well as upon formal documentation. Since the accessibility of such resources, following Kanhabua (2013), may be expected to diminish over time, we expect that the learnability of terms degrades as time passes. We also expect that the fidelity of the learner's understanding of the term may likewise be subject to change, resulting in an increased likelihood of change in the way that the learner chooses to apply terms. If the learner actively makes use of the terminology acquired, this has a relatively high probability of resulting in propagation of acquired semantic term shift to future learners.

The predictions made by this model appear to fit well with intuitions about these two cases, but it is important to stress that the model significantly simplifies events. In particular, we have made the assumption that a learner who is directly exposed to the use of a term by an expert user learns it with perfect fidelity, which we know is not the case. In practice, learning may be a partial or incomplete process, which raises the probability that a variant form of h will be created and come into use. In the event that a variant is created, a large and active community may prove to be *more* prone to propagating the variant, just as they would be more likely to rapidly learn any termset. This is especially true if it proves to be 'fitter' in an evolutionary sense than the original. For example, if a variant fits a group of users' needs better than the original, the variant will be more attractive and hence propagate more rapidly than the original, although this aspect of the model is out of scope for this paper.

4.2 Risk management

A useful outcome of a sustainability model is the ability to power decision-support applications on the individual and organisational level. This model uses observable features of a terminology set in use, notably a combination of community size and level of connectivity, to estimate, in the absence of detailed information, the 'learnability' (in terms of time cost) of a terminology set. The first of the proposed extensions to this model allows the effects of time to be modelled, drawing a distinction between a venerable application profile that is in frequent use and a similarly aged application profile that is in a state of abandonment. The second permits probable fidelity of duplication to be estimated; the practical use of such an approach is likely to depend on validation against real-life datasets. If validated experimentally, however, this model permits us not only to discuss the 'vitality' of a metadata artefact in terms of user count, but also to take into account the effects of periods of disuse and discontinuity in user community. Finally, it also permits us to take into account the likely effects of community structure and size on semantic shift and eventual evolution, where semantic evolution is here defined as propagation of opportunistic or accidental changes that prove to be beneficial to users.

It may with justice be remarked that the likelihood of popular metadata artefacts suffering from temporary abandonment or periods of disuse is low, and this is certainly the case. However, in many domains, especially in the experimental sciences, we find that temporary uptake and use of a metadata standard is a common phenomenon, and is often aligned to the vagaries of funding as well as to trends within the relevant research community. In such cases it is common to see temporarily active 'islands' of usage of specialist standards; understanding the likely outcome of this pattern is useful in understanding how artefacts resulting from such activity may best be understood, preserved and shared.

4.3 Metadata management best practices

Existing best practice in the domain of metadata management handles change (popularly termed evolution) of metadata schemas via an all-or-nothing approach: either a term is deprecated, or it is not; either a term is used, or it is not. Provenance has therefore become extremely significant in DCMI terms as the number of extant records continues to rise, as provenance metadata provides us with useful clues as to the characteristics of each record. Yet with attentive observation of an application domain, it is likely to become possible to actively and explicitly track change, information that can be used to guide further use of schemas and application profiles themselves and to guide our use of the information annotated: it is also a useful resource in mapping change within the application domain itself. For now, many questions remain: how do we gather and store such information? If it were available to us, how might we make use of it in our thinking and practice?

5. Conclusion and further work

In this paper, we have made use of a model inspired by theories of language acquisition to explore the effect of sparse and connected community groupings upon a learner hoping to develop an understanding of the usage of a specialised termset such as an application profile. This model suggests that scenarios involving discontinuity or high rates of change in community membership are more likely to suffer from issues with making use of that application profile. To increase the speed of term acquisition under these circumstances, users will be more likely to make use of lower fidelity learning strategies, including access to documentation, which unless updated becomes less accessible over time, and the use of undocumented exemplars from which to learn. We suggest that these are likely sources for semantic change. Finally, we remark that some occurrences of semantic change may have beneficial effects on the pragmatic usefulness of the termset, and are therefore likely to propagate within the relevant user community when they do occur; hence, while group discontinuity reduces the speed of adoption of termsets, we also expect it to increase the proportional likelihood that semantic evolution occurs; we expect to explore this possibility in future work.

References

- Axelrod, R., & Tesfatsion, L. (2006). Appendix AA Guide for Newcomers to Agent-Based Modeling in the Social Sciences. *Handbook of computational economics*, 2, 1647-1659.
- Baker, T., Dekkers, M., Heery, R., Patel, M., & Salokhe, G. (2001). What terms does your metadata use? Application profiles as machine-understandable narratives. *Journal of Digital information*, 2(2).
- Baruzzo, A., Casoto, P., Dattolo, A., & Tasso, C. (2009). Handling Evolution in Digital Libraries. *IRCDL*, 9, 34-50.
- Berger, J. (1972). *Ways of seeing*. London: BBC.
- Bloomfield, Leonard (1933), *Language*, New York: Allen & Unwin
- Box, George E. P.; Norman R. Draper (1987). *Empirical Model-Building and Response Surfaces*, p. 424, Wiley. ISBN 0471810339.
- Cabot, J., Easterbrook, S., Horkoff, J., Lessard, L., Liaskos, S., & Mazón, J. (2009, May). Integrating sustainability in decision-making processes: A modelling strategy. In *Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on* (pp. 207-210). IEEE.
- Calero, C., Moraga, M. Á., Bertoa, M. F., & Duboc, L. (2015). Quality in Use and Software Greenability.
- Cudré-Mauroux, P., Aberer, K., Abdelmoty, A. I., Catarci, T., Damiani, E., Illaramendi, A., & De Tré, G. (2006). Viewpoints on emergent semantics. In *Journal on Data Semantics VI* (pp. 1-27). Springer Berlin Heidelberg.
- Day, M. (2004). Preservation metadata initiatives: practicality, sustainability, and interoperability.
- Dushay, N., & Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use.
- Franch, X., Susi, A., Annosi, M. C., Ayala, C. P., Glott, R., Gross, D., & Siena, A. (2013). Managing Risk in Open Source Software Adoption. In *ICSOFT* (pp. 258-264).
- Gulla, J. A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., & Cerrato, O. (2010, April). Semantic Drift in Ontologies. In *WEBIST* (2) (pp. 13-20).
- Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, 25(September).
- Messerschmitt, D. G., & Szyperski, C. (2003). *Software ecosystem. Understanding an Indispensable Technology and Industry*. Massachusetts Institute of Technology, Cambridge, MA.
- Kanhubua, N., Niederée, C., & Siberski, W. (2013). Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the 10th International Conference on Preservation of Digital Objects, iPres* (Vol. 2013).
- Kapitzke, C. (2001). Information literacy: The changing library. *Journal of Adolescent and Adult Literacy*, 44(5), 450-456.
- Kaplan, F., Oudeyer, P. Y., & Bergen, B. (2008). Computational models in the debate over language learnability. *Infant and Child Development*, 17(1), 55-80.
- Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge: MIT press.
- Olson, H. A. (1998). Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library trends*, 47(2), 233-254.
- Olson, H. A. (2001). The power to name: Representation in library catalogs. *Signs*, 639-668.
- Olson, H. A. (2002). *The power to name: locating the limits of subject representation in libraries*. Kluwer Academic Pub.
- Penzenstadler, B., Bauer, V., Calero, C., & Franch, X. (2012). Sustainability in software engineering: A systematic literature review.
- Rajaram, T., & Das, A. (2010). Modeling of interactions among sustainability components of an agro-ecosystem using local knowledge through cognitive mapping and fuzzy inference system. *Expert Systems with Applications*, 37(2), 1734-1744.
- Safarzyńska, K., Frenken, K., & van den Bergh, J. C. (2012). Evolutionary theorizing and modeling of sustainability transitions. *Research Policy*, 41(6), 1011-1024.
- Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. *The state of digital preservation: an international perspective*, 4-31.



Metadata Migration—Session 3

Guidance, Please! Towards a Framework for RDF-based Constraint Languages

Thomas Bosch
GESIS – Leibniz Institute
for the Social Sciences, Germany
thomas.bosch@gesis.org

Kai Eckert
Stuttgart Media University, Germany
eckert@hdm-stuttgart.de

Abstract

In the context of the DCMI RDF Application Profile task group and the W3C Data Shapes Working Group solutions for the proper formulation of constraints and validation of RDF data on these constraints are being developed. Several approaches and constraint languages exist but there is no clear favorite and none of the languages is able to meet all requirements raised by data practitioners. To support the work, a comprehensive, community-driven database has been created where case studies, use cases, requirements and solutions are collected. Based on this database, we have hitherto published 81 types of constraints that are required by various stakeholders for data applications. We are using this collection of constraint types to gain a better understanding of the expressiveness of existing solutions and gaps that still need to be filled. Regarding the implementation of constraint languages, we have already proposed to use high-level languages to describe the constraints, but map them to SPARQL queries in order to execute the actual validation; we have demonstrated this approach for the Web Ontology Language in its current version 2 and Description Set Profiles. In this paper, we generalize from the experience of implementing OWL 2 and DSP by introducing an abstraction layer that is able to describe constraints of any constraint type in a way that mappings from high-level constraint languages to this intermediate representation can be created more or less straight-forwardly. We demonstrate that using another layer on top of SPARQL helps to implement validation consistently across constraint languages, simplifies the actual implementation of new languages, and supports the transformation of semantically equivalent constraints across constraint languages.

Keywords: RDF validation; RDF constraints; RDF constraint types, RDF validation requirements; Linked Data; Semantic Web

1. Introduction

The proper validation of RDF data according to constraints is a common requirement of data practitioners. Among the reasons for the success of XML is the possibility to formulate fine-grained constraints to be met by the data and to validate the data according to these constraints using powerful systems like DTD, XML Schema, RELAX NG, or Schematron.

In 2013, the W3C organized the *RDF Validation Workshop*¹ where experts from industry, government, and academia discussed first RDF validation use cases. In 2014, two working groups on RDF validation were established: the *W3C RDF Data Shapes Working Group*² and the *DCMI RDF Application Profiles Task Group*.³ We collected the findings of these working groups and initiated a database of RDF validation requirements⁴ with the intention to collaboratively collect case studies, use cases, requirements, and solutions in a comprehensive and structured way (Bosch & Eckert, 2014a). Based on our work in the DCMI and in cooperation with the W3C

¹ <http://www.w3.org/2012/12/rdf-val/>

² <http://www.w3.org/2014/rds/charter>

³ <http://wiki.dublincore.org/index.php/RDF-Application-Profiles>

⁴ Online available at: <http://purl.org/net/rdf-validation>

working group, we identified by today 81 constraint types, where each type corresponds to a specific requirement in the database. In a technical report, we explain each constraint type in detail and give examples for each represented by different constraint languages (Bosch, Nolle, Acar, & Eckert, 2015).

Various constraint languages exist or are being developed that support more or less of these constraint types. For our work, we focus on the following four as the ones that are most popular among data practitioners, often mentioned on mailing lists and/or being candidates or prototypes for the upcoming W3C recommendation: *Description Set Profiles (DSP)*,⁵ *Resource Shapes (ReSh)*,⁶ *Shape Expressions (ShEx)*,⁷ and the *Web Ontology Language (OWL)*.⁸ Despite the fact that OWL is arguably not a constraint language, it is widely used in practice as such under the closed-world and unique name assumptions.

With its direct support of validation via SPARQL, the *SPARQL Inferencing Notation (SPIN)*⁹ is also very popular to formulate and check constraints (Fürber & Hepp, 2010). We consider SPIN as a low-level language in contrast to the other constraint languages where specific language constructs exist to define constraints in a declarative and in comparison more intuitive way – although SPARQL aficionados might object particularly to the latter point.

The power of SPIN is shown in Table 1, where we list the fraction (and absolute numbers in brackets) of how many constraint types each of these languages supports (Bosch et al., 2015). We further see that OWL 2 is currently the most expressive high-level constraint language, at least according to the pure number of constraint types supported. This does not preclude that other constraint languages are better suited for certain applications, either because they support some types that are not supported by OWL or because the constraint representation is more appealing to the data practitioners – producers as well as consumers who again might have different needs and preferences.

TABLE 1: Constraint Type Specific Expressivity of Constraint Languages

DSP	ReSh	ShEx	OWL 2	SPIN
17.3 (14)	25.9 (21)	29.6 (24)	67.9 (55)	100.0 (81)

We formerly demonstrated that a high-level constraint language like OWL 2 and DSP can be implemented by mapping the language to SPIN using SPARQL CONSTRUCT queries (Bosch & Eckert, 2014b). We provide a validation environment where own mappings from arbitrary constraint languages can be provided and tested.¹⁰ The only limitations are that the constraints have to be expressed in RDF and that the constraint language is expressible in SPARQL.

The constraint type *minimum qualified cardinality restrictions* which corresponds to the requirement *R-75*¹¹ can be instantiated to formulate the constraint that publications must have at least one author which must be a person. This constraint can be expressed as follows using different constraint languages:

⁵ <http://dublincore.org/documents/2008/03/31/dc-dsp/>

⁶ <http://www.w3.org/Submission/2014/SUBM-shapes-20140211/>

⁷ <http://www.w3.org/Submission/2014/SUBM-shex-primer-20140602/>

⁸ <http://www.w3.org/TR/owl2-syntax/>

⁹ <http://spinrdf.org/>

¹⁰ Online available at: <http://purl.org/net/rdfval-demo>, source code online available at: <https://github.com/boschthomas/rdf-validator>.

¹¹ Requirements are identified in the database by an R and a number, additionally an alphanumeric identifier is provided, in this case *R-75-MINIMUM-QUALIFIED-CARDINALITY-ON-PROPERTIES*. Online at: <http://lelystad.informatik.uni-mannheim.de/rdf-validation/?q=node/82>

```

OWL 2: Publication a owl:Restriction ;
      owl:minQualifiedCardinality 1 ;
      owl:onProperty author ;
      owl:onClass Person .

ShEx: Publication { author @Person{1, } }

ReSh: Publication a rs:ResourceShape ; rs:property [
      rs:propertyDefinition author ;
      rs:valueShape Person ;
      rs:occurs rs:One-or-many ; ] .

DSP: [ dsp:resourceClass Publication ; dsp:statementTemplate [
      dsp:minOccur 1 ;
      dsp:property author ;
      dsp:nonLiteralConstraint [ dsp:valueClass Person ] ] ] .

SPIN: CONSTRUCT { [ a spin:ConstraintViolation ... . ] } WHERE {
      ?this
      a ?C1 ;
      ?p ?o .
      BIND ( qualifiedCardinality( ?this, ?p, ?C2 ) AS ?c ) .
      BIND( STRDT ( STR ( ?c ), xsd:nonNegativeInteger ) AS ?cardinality ) .
      FILTER ( ?cardinality < 1 ) .
      FILTER ( ?C1 = Publication ) .
      FILTER ( ?C2 = Person ) .
      FILTER ( ?p = author ) . }

SPIN function qualifiedCardinality:
SELECT ( COUNT ( ?arg1 ) AS ?c ) WHERE { ?arg1 ?arg2 ?o . ?o a ?arg3 . }

```

Note that the SPIN representation of the constraint is *not* a SPIN mapping to implement the constraint, but a direct expression of the constraint using a SPARQL CONSTRUCT query that creates a spin:ConstraintViolation if the constraint is violated.

It can be seen that the higher-level constraint languages are comparatively similar, there seems to be a pattern, a common way to express this type of constraint. Therefore, a mapping from a high-level language to another high-level language would be considerably easier. Unfortunately, there is not (yet) a high-level language that supports all constraint types.

The creation of mappings of constraint languages to SPIN to implement their validation is in many cases not straight-forward and requires profound knowledge of SPARQL, as the following example demonstrates. In this example, the validation of the *minimum qualified cardinality restrictions* constraint type is implemented for DSP:

```

CONSTRUCT {
  _:constraintViolation
    a spin:ConstraintViolation ;
    rdfs:label ?violationMessage ;
    spin:violationRoot ?this ;
    spin:violationPath ?property ;
    spin:violationSource ?violationSource . }
WHERE {
  ?this a ?resourceClass .
  ?descriptionTemplate
    dsp:resourceClass ?resourceClass ;
    dsp:statementTemplate ?statementTemplate .
  ?statementTemplate
    dsp:minOccur ?minimum ;
    dsp:property ?property ;
    dsp:nonLiteralConstraint ?nonLiteralConstraint .
  ?nonLiteralConstraint dsp:valueClass ?valueClass .
  BIND ( qualifiedCardinality( ?this, ?property, ?valueClass ) AS ?cardinality ) .
  FILTER ( ?cardinality < ?minimum ) . }

```

The SPIN mappings for OWL 2 and DSP are rather complicated and can be found in

the mappings provided by us.¹²

In this paper, we build on the experience gained from mapping several constraint languages to SPIN and from the analysis of the identified constraint types to create an intermediate layer, a framework that is able to describe the mechanics of all constraint types and that can be used to map high-level languages more easily.

2. Motivation

Even with an upcoming W3C recommendation, it can be expected that several constraint languages will be used in practice in future – consider the situation in the XML world, where a standardized schema language was available from the beginning and yet additional ways to formulate and check constraints have been created. Therefore, semantically equivalent constraints represented in different languages will exist. This raises two questions:

1. How can we ensure that two semantically equivalent constraints are actually validated consistently?
2. How can we support the transformation of semantically equivalent constraints from one constraint language to another?

Consistent implementation. Even though SPIN provides a convenient way to represent constraints and to validate data according to these constraints, the implementation of a high-level constraint language still requires a tedious mapping to SPIN with a certain degree of freedom as to how a constraint violation is actually represented and how exactly the violation of the constraint is checked. Our framework therefore provides a common ground that is solely based on the abstract definitions of the constraint types, as identified in our database. By providing a SPIN mapping for each constraint type,¹³ it is ensured that the details of the SPIN implementation are consistent irrespective of the constraint language and that the validation leads always to exactly the same results.

Constraint transformation. Consistent implementations of constraint languages provide some advantage, but it could be argued that they are not important enough to justify the additional layer. The situation, however, is different when transformations from one constraint language to another are desired, i.e., to transform a *specific constraint* sc_α of any constraint type expressed by language α into a semantically equivalent *specific constraint* sc_β of the same constraint type represented by any other language β . By defining mappings between equivalent *specific constraints* and the corresponding *generic constraint* (gc) we are able to convert them automatically:

$$gc = m_\alpha(sc_\alpha)$$

$$sc_\beta = m'_\beta(gc)$$

Thereby, we do not need to define mappings for each constraint type and each possible combination of constraint languages. Assuming that we are able to express a single constraint type like *minimum qualified cardinality restrictions* within 10 languages, $n \cdot n - 1 = 90$ mappings would be needed – as mappings generally are not invertible. With an intermediate generic

¹² OWL 2 mapping online available at: https://github.com/boschthomas/rdf-validation/blob/b6a275fb5d71a92ae33d3b6aadd5f447351214b7/SPIN/OWL2_SPIN-Mapping.ttl; DSP mapping online available at: https://github.com/boschthomas/rdf-validation/blob/b6a275fb5d71a92ae33d3b6aadd5f447351214b7/SPIN/DSP_SPIN-Mapping.ttl#L4665

¹³ RDF-CV to SPIN online available at: <https://github.com/boschthomas/RDF-CV-2-SPIN>

representation of constraints, on the other side, we only need to define for each constraint type $2n = 20$ mappings – where 10 mappings should already exist if we have an implementation in our framework. To summarize, if language developers are willing to provide two mappings – forward (m) and backward (m') – to our framework for each supported constraint type, we not only would get the consistent implementation of all languages, it would also be possible to transform semantically equivalent constraints into all constraint languages.

3. Towards a Framework

When we fully implemented OWL 2 and DSP and to some extend other constraint languages using SPARQL as intermediate language (Bosch & Eckert, 2014b), we found that many mappings actually resemble each other; particularly the mappings of the same constraint type in different languages, but also the mappings of different constraint types, though the latter only on a very superficial, structural level. The basic idea of our framework is very simple: we aim at reducing the representation of constraints to the absolute minimum that has to be provided in a mapping to SPIN to implement the validation for constraint types. Consider again our example

```
SPIN: CONSTRUCT { [ a spin:ConstraintViolation ... . ] } WHERE {
    ?this
      a ?C1 ;
      ?p ?o .
    BIND ( qualifiedCardinality( ?this, ?p, ?C2 ) AS ?c ) .
    BIND( STRDT ( STR ( ?c ), xsd:nonNegativeInteger ) AS ?cardinality ) .
    FILTER ( ?cardinality < 1 ) .
    FILTER ( ?C1 = Publication ) .
    FILTER ( ?C2 = Person ) .
    FILTER ( ?p = author ) . }
```

```
SPIN function qualifiedCardinality:
SELECT ( COUNT ( ?arg1 ) AS ?c ) WHERE { ?arg1 ?arg2 ?o . ?o a ?arg3 . }
```

from above for the SPIN representation of a constraint of the type *minimum qualified cardinality restrictions*:

However this SPIN code looks like, all we have to provide to make it work is the desired minimum cardinality (?cardinality), the property to be constrained (?p), the class whose individuals must hold for the constraint (?C1), and the class for which the property should be

```
OWL 2: Publication a owl:Restriction ;
      owl:minQualifiedCardinality 1 ;
      owl:onProperty author ;
      owl:onClass Person .

ShEx: Publication { author @Person{1, } }

ReSh: Publication a rs:ResourceShape ; rs:property [
      rs:propertyDefinition author ;
      rs:valueShape Person ;
      rs:occurs rs:One-or-many ; ] .

DSP: [ dsp:resourceClass Publication ; dsp:statementTemplate [
      dsp:minOccur 1 ;
      dsp:property author ;
      dsp:nonLiteralConstraint [ dsp:valueClass Person ] ] ] .
```

constrained (?C2). All other variables are bound internally. So we could reduce the effort of the mapping by simply providing these four values, which are readily available in all representations of this constraint type:

In further investigation of all kind of constraints and particularly the list of constraint types, we aimed at identifying the building blocks of such constraints to come up with a concise representation of every constraint type.

3.1. Building Blocks

At the core, we use a very simple conceptual model for constraints (see Figure 1), using a small lightweight vocabulary called *RDF Constraints Vocabulary (RDF-CV)*.¹⁴

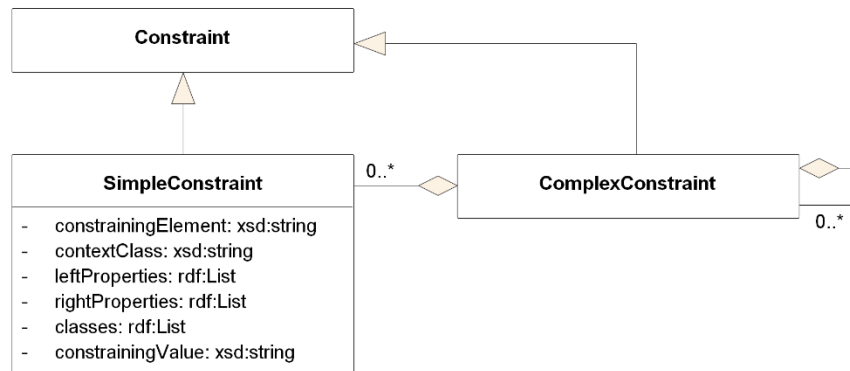


FIG. 1. *RDF Constraints Vocabulary (RDF-CV)* Conceptual Model

RDF constraints are either simple constraints or complex constraints. *Simple constraints* denotes the set of atomic constraints with respect to a single constraining element – we will come to the notion of a constraining element in a second. In contrast, there are *complex constraints*, i.e., the set of constraints which are created out of simple and/or other complex constraints. This structure therefore allows to build complex constraints out of other (simple or complex) constraints. Regarding our database of constraint types, 60% of the constraint types are used to instantiate simple constraints and 26% complex constraints. Constraints of additional 14% of the constraint types are complex constraints as well which can be simplified and therefore formulated as simple constraints if additional constraining elements are introduced to cover them.

The properties describing a simple constraint are very structural, i.e., the properties describe the structure of constraints. The central property is the *constraining element* which refers to one of 103 constraining elements described in our technical report (Bosch et al., 2015). Constraining elements are for example taken from Description Logics, another concrete example would be the SPARQL function REGEX where a regular expression is checked against some property value. In most cases, constraining elements directly correspond to a constraint type, sometimes (as for REGEX) they are shared by several constraint types. Complex constraints again need several constraining elements to be expressed.

Irrespective of and additional to the constraining element, there are properties to describe the actual constraint, they can also be seen as parameters for the constraining element. The *context class* limits the constraint to individuals of a specific class. Depending on the constraining elements, a list of *classes* can be provided, for example to determine the valid classes for a value or to define a class intersection to be used in a constraint. *leftProperties* and *rightProperties* are lists usually containing properties the constraint is applied to. A typical example for a constraint type with a right hand side list of properties would be *literal value comparison (R-43)*, where constraints like `birthDate < deathDate` can be expressed. Finally, the *constraining value* contains a literal value to be checked against; for instance in the case of the REGEX element, it contains the regular expression to be evaluated.

This simple structure plus the constraining elements form the building blocks of our proposed framework. In the technical report (Bosch et al., 2015), we list for every constraint type its representation in our framework which not only shows that constraints of any constraint type can

¹⁴ Formal specification and HTML documentation online available at: <https://github.com/boschthomas/RDF-Constraints-Vocabulary>

indeed be described generically in this way, but which also forms the starting point for any mappings using this framework.

Formal approach and semantics. A cornerstone of the framework is the generic representation of a constraint, which can often be done using Description Logics. For example the *minimum qualified cardinality restriction* can be expressed as $\text{Publication} \sqsubseteq \geq 1 \text{ author.Person}$. This way, the knowledge representation formalism *Description Logics (DL)* (Krötzsch, Simancík, & Horrocks, 2012; Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003; Baader & Nutt, 2003) with its well-studied theoretical properties provides the foundational basis for the framework.

It turned out that 64% of the 81 constraint types are actually expressible in DL. Only for the remaining 36%, other means, i.e., other constraining elements, had to be identified. This is not surprising if we consider that OWL is based on DL. When we talk about using DL to represent constraints, we have to establish once more that the semantics of OWL and DL differ from the semantics of constraint languages regarding the open world assumption (OWA) and the non-unique name assumption (nUNA). Both are usually assumed when dealing with OWL or DL, whereas validation usually assumes a closed world (CWA) and unique naming (UNA), i.e., if a desired property is missing, this leads to a violation and if two resources are named differently, they are assumed to be different resources.

We won't get into details about these assumptions here, but it has to be noted that the applied semantics have to be defined if validation is performed, as the results would differ under different semantics. Precisely, we found that for 56.8% of the constraint types validation results differ if the CWA or the OWA is assumed and for 66.6% of the constraint types validation results are different in case the UNA or the nUNA is assumed (Bosch et al., 2015).

For the purpose of a consistent implementation and transformation of constraints, constraints are considered *semantically equivalent* if they detect the same set of violations regardless of RDF data, which means whenever the constraints are applied to any RDF data they point out the same violations.

3.2. Simple Constraints

In this and the following section, we provide examples for the representation of constraint types within the framework.

The *minimum qualified cardinality restriction (R-75)* $\text{Publication} \sqsubseteq \geq 1 \text{ author.Person}$, which restricts publications to have at least one author which must be a person, is an example of a simple constraint on *author* which holds for all individuals of the class *Publication*. Table 2 displays how the simple constraint is generically represented using the RDF-CV.

TABLE 2: Minimum Qualified Cardinality Restriction as Property Constraint

context class	left property list	right p. list	classes	constraining element	c. value
Publication	author	-	Person	\geq	1

The *constraining element* is an intuitive term which indicates the actual type of constraint. For the majority of the constraint types, there is exactly one constraining element, for instance *property domain (R-25, R-26)* restricts domains of properties and there is only one constraining element with exactly the same identifier *property domain*. Some constraint types, however, need several constraining elements to be expressed, for instance *language tag cardinality (R-48, R-49)* is used to restrict data properties to have a minimum, maximum, or exact number of relationships to literals with selected language tags. Thus, three constraining elements are needed to express each possible constraint of that constraint type. This example also illustrates that the granularity of the constraint types varies and certainly often is debatable. Keep in mind that they correspond

to requirements as identified by the working groups. The constraining elements, as in this example, are closer to atomic elements of constraints.

If constraint types are expressible in DL, associated constraining elements are formally based on DL constructs like concept and role constructors (\sqsubseteq , \sqsupseteq , \sqcap , \sqcup , \neg , \exists , \forall , \geq , \leq), equality ($=$), and inequality (\neq). In case constraint types cannot be expressed in DL such as *data property facets* (R-46) or *literal pattern matching* (R-44), we reuse widely known terms from SPARQL (e.g., REGEX) or XML Schema constraining facets (e.g., *xsd:minInclusive*) as constraining elements. We provide a complete list of all 103 constraining elements which can be used to express constraints of any constraint type (Bosch et al., 2015).

Additional to the constraining element, there are properties of simple constraints which can be seen as parameters for the constraining element. In some cases, a simple constraint is only complete when a *constraining value* is stated in conjunction with the constraining element. Depending on the constraining element, a list of *classes* can be provided, for example to determine the valid classes for a value. The constraining element of the constraint `Publication $\sqsubseteq \geq 1$ author.Person`, e.g., is \geq , the constraining value is *1*, and the list of classes includes the class *Person* which restricts the objects of the property *author* to be persons. The assignment of properties to the left or right property lists depends on the constraining element.

Object property paths (R-55) ensure that if an individual *x* is connected by a sequence of object properties with an individual *y*, then *x* is also related to *y* by a particular object property. As *Stephen-Hawking* is the author of the book *A-Brief-History-Of-Time* whose genre is *Popular-Science*, the object property path `authorOf \circ genre \sqsubseteq authorOfGenre` infers that *Stephen-Hawking* is an author of the genre *Popular-Science*. Thus, when representing the constraint using the RDF-CV (see Table 3), the properties *authorOf* and *genre* are placed on the left side of the constraining element *property path* and the property *authorOfGenre* on its right side. The *context class* limits the constraint to individuals of a specific class. A context class may be an *rdfs:Class*, an *owl:Class* (as sub-class of *rdfs:Class*), or an *rdfs:Datatype* which is both an instance of and a sub-class of *rdfs:Class*. As the *property path* constraint holds for all individuals within the data, the context class is set to the *DL top concept* \top which stands for the super-class of all possible classes.

TABLE 3: Object Property Paths as Property Constraint

context class	left p. list	right p. list	classes	c. element	c. value
\top	authorOf, genre	authorOfGenre	\top	property path	-

Constraints of 36% of the constraint types are not expressible in DL but can still be described using the RDF-CV such as constraints of the type *literal pattern matching* (R-44) which restrict literals to match given patterns. The *universal quantification* (R-91) `Book $\sqsubseteq \forall$ identifier.ISBN` ensures that books can only have valid *ISBN* identifiers, i.e., strings that match a given regular expression.

Even though constraints of the type *literal pattern matching* cannot be expressed in DL, OWL

```
ISBN a RDFS:Datatype ; owl:equivalentClass [ a RDFS:Datatype ;
  owl:onDatatype xsd:string ;
  owl:withRestrictions ( [ xsd:pattern "\d{9}[\d|X]$" ] ) ] .
```

2 can be used to formulate this constraint:

The first OWL 2 axiom explicitly declares *ISBN* to be a datatype. The second OWL 2 axiom defines *ISBN* as an abbreviation for a datatype restriction on *xsd:string*. The datatype *ISBN* can be used just like any other datatype like in the universal quantification above.

Table 4 presents (1) how the not in DL expressible literal pattern matching constraint and (2) how the in DL expressible universal quantification are both represented using the RDF-CV. Thereby, the context class *ISBN*, whose instances must satisfy the literal pattern matching constraint, is reused within the list of classes the universal quantification refers to. The literal pattern matching constraint type introduces the constraining element *REGEX* whose validation has to be implemented once like for any other constraining element.

TABLE 4: Simple Constraints which are not Expressible in DL

context class	left p. list	right p. list	classes	c. element	c. value
ISBN	-	-	xsd:string	REGEX	"^\\d{9}[\\d x]\$"
Book	identifier	-	ISBN	universal quantification	-

3.3. Complex Constraints

Complex constraints of the constraint type *context-specific exclusive or of property groups (R-13)* restrict individuals of given classes to have all properties of exactly one of multiple mutually exclusive property groups. Publications, e.g., are either identified by an ISBN and a title (for

```
Publication {
  ( isbn string , title string ) |
  ( issn string , title string ) }
```

books) or by an ISSN and a title (for periodical publications), but it should not be possible to assign both identifiers to a given publication. This complex constraint is expressible in ShEx:

If *The-Great-Gatsby* is a publication with an ISBN and a title without an ISSN, *The-Great-Gatsby* is considered as a valid publication. This complex constraint is generically expressible in DL:

$$\begin{aligned} \text{Publication} &\sqsubseteq (\neg E \sqcap F) \sqcup (E \sqcap \neg F), \quad E \equiv A \sqcap B, \quad F \equiv C \sqcap D \\ A &\sqsubseteq \geq 1 \text{ isbn.string} \sqcap \leq 1 \text{ isbn.string}, \quad B \sqsubseteq \geq 1 \text{ title.string} \sqcap \leq 1 \text{ title.string} \\ C &\sqsubseteq \geq 1 \text{ issn.string} \sqcap \leq 1 \text{ issn.string}, \quad D \sqsubseteq \geq 1 \text{ title.string} \sqcap \leq 1 \text{ title.string} \end{aligned}$$

The DL statements demonstrate that the complex constraint is composed of many other complex constraints (*minimum (R-75)* and *maximum qualified cardinality restrictions (R-76)*) and simple constraints (*intersection (R-15/16)*, *disjunction (R-17/18)*, and *negation (R-19/20)*). Constraints of almost 14% of the constraint types are complex constraints which can be simplified and therefore formulated as simple constraints when using them in terms of syntactic sugar. As *exact (un)qualified cardinality restrictions (R-74/80) (=n)* and *exclusive or of property groups (R-13)* are constraint types of frequently used complex constraints, we propose to simplify them in form of simple constraints. As a consequence, the *context-specific exclusive or of property groups* complex constraint is represented as a generic constraint by means of the RDF-CV more intuitively and concisely (see Table 5).

TABLE 5: Simplified Complex Constraints

context class	left p. list	right p. list	classes	c. element	c. value
Publication	-	-	E, F	exclusive or	-
E	-	-	A, B	intersection	-
F	-	-	C, D	intersection	-

A	isbn	-	string	=	1
B	title	-	string	=	1
C	issn	-	string	=	1
D	title	-	string	=	1

The *primary key properties* (R-226) constraint type is often useful to declare a given (datatype) property as the primary key of a class, so that a system can enforce uniqueness. Books, e.g., are uniquely identified by their ISBN, i.e., the property *isbn* is inverse functional ($\text{funct } \text{isbn}^-$) which can be represented using the RDF-CV in form of a complex constraint consisting of two simple constraints (see Table 6). The meaning of these simple constraints is that ISBN identifiers can only have isbn^- relations to at most one distinct book.

TABLE 6: Primary Key Properties as Complex Constraints

context class	left p. list	right p. list	classes	c. element	c. value
T	isbn ⁻	isbn	-	inverse property	-
Book	isbn ⁻	-	-	≤	1

Keys, however, are even more general, i.e., a generalization of inverse functional properties (Schneider, 2009). A key can be a datatype, an object property, or a chain of properties. For these generalization purposes, as there are different sorts of keys, and as keys can lead to undecidability, DL is extended with a special construct *keyfor* (Lutz, Areces, Horrocks, & Sattler, 2005). When using *keyfor* (*isbn keyfor Book*), the complex constraint can be simplified and thus

```
[
  a rdfcv:SimpleConstraint ;
  rdfcv:contextClass Book ;
  rdfcv:leftProperties ( isbn ) ;
  rdfcv:constrainingElement "primary key" ] .
```

formulated as a simple constraint which looks like the following in concrete RDF turtle syntax:

Complex constraints of frequently used constraint types which correspond to DL axioms like *transitivity*, *symmetry*, *asymmetry*, *reflexivity* and *irreflexivity* can also be simplified in form of simple constraints. Although these DL axioms are expressible by basic DL features, they can also be used in terms of syntactic sugar.

Constraints of the *irreflexive object properties* (R-60) constraint type ensure that no individual is connected by a given object property to itself (Krötzsch et al., 2012). With the irreflexive object property constraint $T \sqsubseteq \neg \exists \text{authorOf.Self}$, e.g., one can state that individuals cannot be authors of themselves. When represented using the RDF-CV, the complex constraint aggregates three simple constraints (see Table 7).

TABLE 7: Irreflexive Object Properties as Complex Constraints

context class	left p. list	right p. list	classes	c. element	c. value
$\exists \text{authorOf.Self}$	authorOf	-	Self	existential quantification	-
$\neg \exists \text{authorOf.Self}$	-	-	$\exists \text{authorOf.Self}$	negation	-
T	-	-	T, $\neg \exists \text{authorOf.Self}$	sub-class	-

When using the *irreflexive object property* constraint in terms of syntactic sugar, the complex constraint can be expressed more concisely in form of a simple property constraint with exactly the same semantics (see Table 8):

TABLE 8: Irreflexive Object Properties as Simple Constraints

context class	left p. list	right p. list	classes	c. element	c. value
T	authorOf	-	-	irreflexive property	-

3.4. Mapping Implementation

Using the framework for the implementation of a constraint language is straight-forward. For each language construct, the corresponding constraint type has to be identified. Again we use the constraint `Publication $\sqsubseteq \geq 1$ author.Person` of the type *minimum qualified cardinality restrictions* (R-75) which is supported in OWL 2:

```
:Publication
  a owl:Restriction ;
  owl:minQualifiedCardinality 1 ;
  owl:onProperty :author ;
  owl:onClass :Person .
```

From Table 2, we know the representation in our framework, which corresponds to the following RDF representation using the RDF-CV:

```
[
  a rdfcv:SimpleConstraint ;
  rdfcv:contextClass :Publication ;
  rdfcv:leftProperties ( :author ) ;
  rdfcv:classes ( :Person ) ;
  rdfcv:constrainingElement "minimum qualified cardinality restriction" ;
  rdfcv:constrainingValue 1 ] .
```

The mapping simply constructs this generic representation out of the specific OWL 2 representation using a SPARQL CONSTRUCT query:

```
owl:Thing
  spin:rule [ a sp:Construct ; sp:text ""
    CONSTRUCT {
      :minimum-qualified-cardinality-restrictions
        a rdfcv:SimpleConstraint ;
        rdfcv:contextClass ?this ;
        rdfcv:leftProperties :leftProperties ;
        rdfcv:classes :classes ;
        rdfcv:constrainingElement "minimum qualified cardinality restriction"
      ;
        rdfcv:constrainingValue ?cv .
      :leftProperties
        rdf:first ?lpl ;
        rdf:rest rdf:nil .
      :classes
        rdf:first ?cl ;
        rdf:rest rdf:nil . }
    WHERE {
      ?this
        a owl:Restriction ;
        owl:minQualifiedCardinality ?cv ;
        owl:onProperty ?lpl ;
        owl:onClass ?cl . } "" ; ] .
```

The SPIN engine is used to execute the mapping, the property *spin:rule* links an `rdfs:Class` with SPARQL CONSTRUCT queries. Each query defines an inference rule that is applied to all instances of the associated class and its subclasses. The inference rule defines how additional

triples can be inferred from what is stated in the WHERE clause. For each binding of the pattern in the WHERE clause of the rule, the triple templates from the CONSTRUCT clause are instantiated and added as inferred triples to the underlying model. At query execution time, the SPARQL variable *?this* is bound to the current instance of the class. As each resource per default is assigned to the class *owl:Thing*, this inference rule is evaluated for each subject of the input RDF graph.

The framework and therefore the constraint types are implemented in exactly the same way by providing other SPIN mappings which encompass the SPIN/SPARQL queries that validate constraints and produce constraint violation messages if a constraint is violated, as described in our previous paper about the DSP implementation (Bosch & Eckert, 2014b).¹⁵

3.5. Constraint Transformation

As stated in Section 2, we see a huge potential in the possibility to transform semantically equivalent constraints from one high-level constraint language to another via the RDF-CV representation, to avoid that every possible combination of constraint languages has to be mapped

```

owl:Thing
  spin:rule [ a sp:Construct ; sp:text """
    CONSTRUCT {
      ?cc
      a owl:Restriction ;
      owl:minQualifiedCardinality ?cv ;
      owl:onProperty ?lp1 ;
      owl:onClass ?c1 . }

    WHERE {
      ?this
      a rdfcv:SimpleConstraint ;
      rdfcv:contextClass ?cc ;
      rdfcv:leftProperties ?leftProperties ;
      rdfcv:classes ?classes ;
      rdfcv:constrainingElement "minimum qualified cardinality restriction"
    ;
      rdfcv:constrainingValue ?cv .
      ?leftProperties
      rdf:first ?lp1 ;
      rdf:rest rdf:nil .
      ?classes
      rdf:first ?c1 ;
      rdf:rest rdf:nil . } """ ; ] .

```

separately. The following SPIN inference rule exemplifies this approach and provides a mapping from RDF-CV back to the OWL 2 constraint of the type *minimum qualified cardinality restrictions*:

It can be seen that the mapping is quite similar to the first mapping and basically simply switches the CONSTRUCT and WHERE part of the query, with slight adjustment in the structure of the variables. Potentially an even simpler representation for the mapping could be found that would enable the creation of forward and backward mappings out of it. We didn't investigate this further, though, and it is not yet clear if there can be cases where the backward mapping is more different.

4. Related Work

In this section, we present current languages for RDF constraint formulation and RDF data validation. SPIN, SPARQL, OWL 2, ShEx, ReSh, and DSP are the six most promising and

¹⁵ At the time of this writing, not all mappings for the constraint types are implemented, but of course the implementations can be complemented and adapted to own requirements, as needed. The most recent implementation can be found here: <https://github.com/boschthomas/rdf-validation/blob/master/SPIN/RDF-CV-2-SPIN.ttl>

mostly used constraint languages. In addition, the W3C Data Shapes Working Group currently develops SHACL, an RDF vocabulary for describing RDF graph structures.

The *SPARQL Query Language for RDF* (Harris & Seaborne, 2013) is generally seen as the method of choice to validate RDF data according to certain constraints (Fürber & Hepp, 2010), although, it is not ideal for their formulation. In contrast, high-level constraint languages are comparatively easy to understand and constraints can be formulated more concisely. Declarative languages may be placed on top of SPARQL and SPIN when using them as implementation languages. The *SPARQL Inferencing Notation (SPIN)*¹⁶ (Knublauch, Hendler, & Idehen, 2011) provides a vocabulary to represent SPARQL queries as RDF triples and uses SPARQL to specify logical constraints and inference rules (Fürber & Hepp, 2010). Kontokostas et al. define 17 data quality integrity constraints represented as SPARQL query templates called *Data Quality Test Patterns (DQTP)* (Kontokostas et al., 2014).

The *Web Ontology Language (OWL)* (Hitzler, Krötzsch, Parsia, Patel-Schneider, & Rudolph, 2012) formally specifies the intended semantics of conceptual models about data and therefore enables software to understand data. OWL has become a popular standard for data representation, data exchange, and data integration of heterogeneous data sources. Besides that, the retrieval of data benefits from semantic knowledge specified using OWL. In combination with the OWL-based *Semantic Web Rule Language (SWRL)* (Horrocks et al., 2004), OWL provides facilities for developing very powerful reasoning services. Reasoning on RDF data enables to derive implicit data out of explicitly stated data. OWL is based on formal logic and on the subject-predicate-object triples from RDF. OWL is actually a description logic with underlying formal semantics which allows one to assign truth values to syntactic expressions. OWL specifies semantic information about specific domains, describes relations between domain classes, and thus allows the sharing of conceptualizations.

Because of the design of OWL for reasoning, there are claims that OWL cannot be used for validation. In practice, however, OWL is well-spread and RDFS/OWL constructs are widely used to tell people and applications about how valid instances should look like. In general, RDF documents follow the syntactic structure and the semantics of RDFS/OWL ontologies which could therefore not only be used for reasoning but also for validation.

Stardog Integrity Constraint Validation (ICV) and the *Pellet Integrity Constraint Validator (ICV)* use OWL 2 constructs to formulate constraints. The Pellet ICV¹⁷ is a proof-of-concept extension for the OWL 2 DL reasoner *Pellet* (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007). Stardog ICV¹⁸ validates RDF data stored in a Stardog database according to constraints which may be written in SPARQL, OWL 2, or SWRL (Horrocks et al., 2004).

Shape Expressions (ShEx) (Prud'hommeaux, 2014; Solbrig & Prud'hommeaux, 2014; Prud'hommeaux, Labra Gayo, & Solbrig, 2014; Boneva et al., 2014) specifies a language whose syntax and semantics are similar to regular expressions. ShEx associate RDF graphs with labeled patterns called *shapes* which are used to express formal constraints on the content of RDF graphs. *Resource Shapes (ReSh)* (A. Ryman, 2014) defines its own vocabulary for specifying shapes of RDF resources. Ryman, Hors, and Speicher define *shape* as a description of the set of triples a resource is expected to contain and as a description of the integrity constraints those triples are required to satisfy (A. G. Ryman, Hors, & Speicher, 2013).

The *Dublin Core Application Profile (DCAP)* and *Bibframe Profiles* are approaches to specify profiles for application-specific purposes. The term *profile* is widely used to refer to a document that describes how standards or specifications are deployed to support the requirements of a particular application, function, community, or context. In the metadata community, the term *application profile* has been applied to describe the tailoring of standards for specific

¹⁶ <http://spinrdf.org>

¹⁷ <http://clarkparsia.com/pellet/icv>

¹⁸ http://docs.stardog.com/#_validating_constraints

applications. A *Dublin Core Application Profile (DCAP)* (Coyle & Baker, 2009) defines metadata records which meet specific application needs while providing semantic interoperability with other applications on the basis of globally defined vocabularies and models. The *Singapore Framework for Dublin Core Application Profiles* (Nilsson, Baker, & Johnston, 2008) is a framework for designing metadata and for defining DCAPs. The framework comprises descriptive components that are necessary or useful for documenting DCAPs.

The *DCMI Abstract Model* (Powell, Nilsson, Naeve, Johnston, & Baker, 2007) is required for formalizing a notion of machine-processable application profiles. It specifies an abstract model for Dublin Core metadata which is independent of any particular encoding syntax. Its primary purpose is to specify the components used in Dublin Core metadata. Nilsson et al. (Nilsson, Powell, Johnston, & Naeve, 2008) depict how the constructs of the DCMI Abstract Model are represented using the abstract syntax of the RDF model. A *Description Set Profile (DSP)* (Nilsson, 2008) is a generic constraint language which is used to formally specify structural constraints on sets of resource descriptions within an application profile. DSP constrains resources that may be described by descriptions in a description set, the properties that may be used, and the values properties may point to. *BIBFRAME*¹⁹ (Kroeger, 2013; Godby, Carol Jean and Denenberg, Ray, 2015; Miller, Eric and Ogbuji, Uche and Mueller, Victoria and MacDougall, Kathy, 2012) is the result of the *Bibliographic Framework Initiative* and defines a vocabulary (Library of Congress, 2014a, 2014c) which has a strong overlap with DSP. *BIBFRAME Profiles* (Library of Congress, 2014b) are essentially identical to DCAPs.

*Schemarama*²⁰ is a validation technique for specifying the types of sub-graphs you want to have connected to a particular set of nodes in an RDF Graph. Schemarama allows to check that RDF data has required properties. Schemarama is based on Schematron (ISO/IEC, 2006), an XML schema and XML structure validation language which works by finding tree patterns within an XML document. Schemarama is also based on the *Squish RDF Query language* (Miller, 2001), an SQL-like query language for RDF, instead of SPARQL.

In addition to the formulation of constraints, SPIN (open source API), Stardog ICV (as part of the Stardog RDF database), DQTP (tests), Pellet ICV (extension of Pellet OWL 2 DL reasoner) and ShEx offer executable validation systems using SPARQL as implementation language.

The W3C Data Shapes Working Group currently develops *SHACL* (Knublauch, 2015; Boneva & Prud'hommeaux, 2015; Prud'hommeaux, 2015), the *Shapes Constraint Language*, an RDF vocabulary for describing RDF graph structures. Some of these graph structures are captured as *shapes*, which group together constraints about the same RDF nodes. Shapes provide a high-level vocabulary to identify predicates and their associated cardinalities, datatypes and other constraints. Additional constraints can be associated with shapes using SPARQL and similar executable languages. These executable languages can also be used to define new high-level vocabulary terms. SHACL shapes can be used to communicate data structures associated with some process or interface, generate or validate data, or drive user interfaces.

5. Conclusion and Future Work

In this paper, we outlined our idea of a general framework to support the mapping of high-level constraint languages to a generic representation, which can directly be validated by providing a mapping from the generic representation to SPIN/SPARQL queries to actually validate data against constraints provided in the high-level language. The framework consists of a very simple conceptual model using the *RDF Constraints Vocabulary (RDF-CV)* which has been introduced in this paper. The core of the framework is the definition of 103 constraining elements that are used to define constraints of all 81 constraint types that to date have been identified within the DCMI RDF Application Profiles Task Group and in cooperation with the W3C Data Shapes

¹⁹ <http://bibframe.org>

²⁰ <http://www.xml.com/pub/a/2001/02/07/schemarama.html>

Working Group. The full definition of all constraint types and the generic representation of the types in RDF-CV is provided in an accompanying technical report (Bosch et al., 2015).

We have demonstrated how the framework can be used to map a constraint language to RDF-CV and also how to map back from RDF-CV to the constraint language. The latter enables the transformation of semantically equivalent constraints from one constraint language to another via the RDF-CV intermediate representation.

We think that this approach is suitable

1. to implement the validation of constraints consistently across constraint languages,
2. to support the extension of constraint languages when additional constraint types should be supported by means of a simple mapping, and
3. to enhance or rather establish the interoperability of different constraint languages.

It is part of future work to finalize the implementation of all 81 constraint types in our *RDF Validator*, to fully map constraint languages to RDF-CV (first and foremost DSP and OWL 2) and of course keep the framework in sync with the ongoing work in the working groups.

References

- Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (Eds.). (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. New York, NY, USA: Cambridge University Press.
- Baader, F., & Nutt, W. (2003). *The Description Logic Handbook*. In F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, & P. F. Patel-Schneider (Eds.), *Basic Description Logics* (pp. 43–95). New York, NY, USA: Cambridge University Press. Retrieved from <http://dl.acm.org/citation.cfm?id=885746.885749>
- Boneva, I., Gayo, J. E. L., Hym, S., Prud'hommeau, E. G., Solbrig, H. R., & Staworko, S. (2014). *Validating RDF with Shape Expressions*. Computing Research Repository (CoRR), abs/1404.1270. Retrieved from <http://arxiv.org/abs/1404.1270>
- Boneva, I., & Prud'hommeaux, E. (2015, July). *Core SHACL Semantics* (W3C Editor's Draft). W3C. (<http://w3c.github.io/data-shapes/semantics/>)
- Bosch, T., & Eckert, K. (2014a). *Requirements on RDF Constraint Formulation and Validation*. In *Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications (DC 2014)*. Austin, Texas, USA. (<http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/257>)
- Bosch, T., & Eckert, K. (2014b). *Towards Description Set Profiles for RDF using SPARQL as Intermediate Language*. In *Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications (DC 2014)*. Austin, Texas, USA. (<http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/270>)
- Bosch, T., Nolle, A., Acar, E., & Eckert, K. (2015). *RDF Validation Requirements - Evaluation and Logical Underpinning*. Computing Research Repository (CoRR), abs/1501.03933. (<http://arxiv.org/abs/1501.03933>)
- Coyle, K., & Baker, T. (2009, May). *Guidelines for Dublin Core Application Profiles* (DCMI Recommended Resource). Dublin Core Metadata Initiative (DCMI). Retrieved from <http://dublincore.org/documents/2009/05/18/profile-guidelines/> (<http://dublincore.org/documents/2009/05/18/profile-guidelines/>)
- Fürber, C., & Hepp, M. (2010). *Using SPARQL and SPIN for Data Quality Management on the Semantic Web*. In W. Abramowicz & R. Tolksdorf (Eds.), *Business Information Systems* (Vol. 47, pp. 35–46). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-12814-1_4 doi: 10.1007/978-3-642-12814-1_4
- Godby, Carol Jean and Denenberg, Ray. (2015, January). *Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC* (Tech. Rep.). Library of Congress. Retrieved from <http://www.oclc.org/research/publications/2015/oclcresearch-loc-linked-data-2015.html> (<http://www.oclc.org/research/publications/2015/oclcresearch-loc-linked-data-2015.html>)
- Harris, S., & Seaborne, A. (2013, March). *SPARQL 1.1 Query Language* (W3C Recommendation). W3C. Retrieved from <http://www.w3.org/TR/2013/REC-sparql11-query-20130321/> (<http://www.w3.org/TR/2013/REC-sparql11-query-20130321/>)
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012, December). *OWL 2 Web Ontology Language Primer* (Second Edition) (W3C Recommendation). W3C. Retrieved from <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/> (<http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>)

- Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosz, B., & Dean, M. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML (W3C Member Submission). W3C. W3C Member Submission. Retrieved from <http://www.w3.org/Submission/SWRL> (<http://www.w3.org/Submission/SWRL>)
- ISO/IEC. (2006, June). ISO/IEC 19757-3:2006 - Information Technology — Document Schema Definition Languages (DSDL) - Part 3: Rule-Based Validation - Schematron (ISO/IEC Specification). ISO/IEC. ([http://standards.iso.org/ittf/PubliclyAvailableStandards/c040833 ISO IEC 19757-3 2006\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c040833%20ISO%20IEC%2019757-3%202006(E).zip))
- Knublauch, H. (2015, July). Shapes Constraint Language (SHACL) (W3C Editor's Draft). W3C. (<http://w3c.github.io/data-shapes/shacl/>) Knublauch, H., Hendler, J. A., & Idehen, K. (2011, February). SPIN - Overview and Motivation (W3C Member Submission). W3C. (<http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/>)
- Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., & Zaveri, A. (2014). Test-driven Evaluation of Linked Data Quality. In Proceedings of the 23rd International World Wide Web Conference (WWW 2014) (pp. 747–758). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from <http://dx.doi.org/10.1145/2566486.2568002> doi: 10.1145/2566486.2568002
- Kroeger, A. (2013). The Road to BIBFRAME: The Evolution of the Idea of Bibliographic Transition into a Post-MARC Future. *Cataloging & Classification Quarterly*, 51(8), 873-890. Retrieved from <http://dx.doi.org/10.1080/01639374.2013.823584> doi:10.1080/01639374.2013.823584
- Krötzsch, M., Simančík, F., & Horrocks, I. (2012). A Description Logic Primer. In J. Lehmann & J. Völker (Eds.), *Perspectives on Ontology Learning*. IOS Press. Library of Congress. (2014a, April). BIBFRAME Authorities (Library of Congress Draft Specification). Library of Congress. Retrieved from <http://www.loc.gov/bibframe/docs/bibframe-authorities.html> (<http://www.loc.gov/bibframe/docs/bibframe-authorities.html>)
- Library of Congress. (2014b, May). BIBFRAME Profiles: Introduction and Specification (Library of Congress Draft). Library of Congress. Retrieved from <http://www.loc.gov/bibframe/docs/bibframe-profiles.html> (<http://www.loc.gov/bibframe/docs/bibframe-profiles.html>)
- Library of Congress. (2014c, April). BIBFRAME Relationships (Library of Congress Draft Specification). Library of Congress. Retrieved from <http://www.loc.gov/bibframe/docs/bibframe-relationships.html> (<http://www.loc.gov/bibframe/docs/bibframe-relationships.html>)
- Lutz, C., Areces, C., Horrocks, I., & Sattler, U. (2005, June). Keys, Nominals, and Concrete Domains. *Journal of Artificial Intelligence Research*, 23(1), 667–726. (<http://dl.ac.org/citation.cfm?id=1622503.1622518>)
- Miller, L. (2001, February). RDF Squish Query Language and Java Implementation (Draft). Retrieved from <http://ilrt.org/discovery/2001/02/squish/> (<http://ilrt.org/discovery/2001/02/squish/>)
- Miller, Eric and Ogbuji, Uche and Mueller, Victoria and MacDougall, Kathy. (2012, November). Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services (Tech. Rep.). Washington, DC, USA: Library of Congress. Retrieved from <http://www.loc.gov/bibframe/pdf/marcl-d-report-11-21-2012.pdf> (<http://www.loc.gov/bibframe/pdf/marcl-d-report-11-21-2012.pdf>)
- Nilsson, M. (2008, March). Description Set Profiles: A Constraint Language for Dublin Core Application Profiles (DCMI Working Draft). Dublin Core Metadata Initiative (DCMI). Retrieved from <http://dublincore.org/documents/2008/03/31/dc-dsp/> (<http://dublincore.org/documents/2008/03/31/dc-dsp/>)
- Nilsson, M., Baker, T., & Johnston, P. (2008, January). The Singapore Framework for Dublin Core Application Profiles (DCMI Recommended Resource). Dublin Core Metadata Initiative (DCMI). Retrieved from <http://dublincore.org/documents/2008/01/14/singapore-framework/> (<http://dublincore.org/documents/2008/01/14/singapore-framework/>)
- Nilsson, M., Powel, A., Johnston, P., & Naeve, A. (2008, January). Expressing Dublin Core Metadata using the Resource Description Framework (RDF) (DCMI Recommendation). Dublin Core Metadata Initiative (DCMI). Retrieved from <http://dublincore.org/documents/2008/01/14/dc-rdf/> (<http://dublincore.org/documents/2008/01/14/dc-rdf/>)
- Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007, June). DCMI Abstract Model (DCMI Recommendation). Dublin Core Metadata Initiative (DCMI). Retrieved from <http://dublincore.org/documents/2007/06/04/abstract-model/> (<http://dublincore.org/documents/2007/06/04/abstract-model/>)
- Prud'hommeaux, E. (2014, June). Shape Expressions 1.0 Primer (W3C Member Submission). W3C. Retrieved from <http://www.w3.org/Submission/2014/SUBM-shex-primer-20140602/> (<http://www.w3.org/Submission/2014/SUBM-shex-primer-20140602/>)
- Prud'hommeaux, E. (2015, July). SHACL-SPARQL (W3C Editor's Draft). W3C. (<http://w3c.github.io/data-shapes/semantics/SPARQL>)
- Prud'hommeaux, E., Labra Gayo, J. E., & Solbrig, H. (2014). Shape Expressions: An RDF Validation and Transformation Language. In Proceedings of the 10th International Conference on Semantic Systems

- (SEMANTiCS) (pp. 32–40). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2660517.2660523> doi: 10.1145/2660517.2660523
- Ryman, A. (2014, February). Resource Shape 2.0 (W3C Member Submission). W3C. Retrieved from <http://www.w3.org/Submission/2014/SUBM-shapes-20140211/> (<http://www.w3.org/Submission/2014/SUBM-shapes-20140211/>)
- Ryman, A. G., Hors, A. L., & Speicher, S. (2013). OSLC Resource Shape: A Language for Defining Constraints on Linked Data. In C. Bizer, T. Heath, T. Berners-Lee, M. Hausenblas, & S. Auer (Eds.), *Proceedings of the International World Wide Web Conference (WWW), Workshop on Linked Data on the Web (LDOW)* (Vol. 996). CEUR-WS.org. Retrieved from <http://dblp.uni-trier.de/db/conf/www/ldow2013.html#RymanHS13>
- Schneider, M. (2009, October). OWL 2 Web Ontology Language RDF-Based Semantics (W3C Recommendation). W3C. (<http://www.w3.org/TR/2009/REC-owl2-rdf-based-semantics-20091027/>)
- Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2), 51–53.
- Solbrig, H., & Prud'hommeaux, E. (2014, June). Shape Expressions 1.0 Definition (W3C Member Submission). W3C. Retrieved from <http://www.w3.org/Submission/2014/SUBM-shex-defn-20140602/> (<http://www.w3.org/Submission/2014/SUBM-shex-defn-20140602/>)

Metadata Quality Control for Content Migration: The Metadata Migration Project at the University of Houston Libraries

Andrew Weidner
University of Houston, USA
ajweidner@uh.edu

Annie Wu
University of Houston, USA
awu@uh.edu

Abstract

The decision to migrate digital objects from one digital asset management system to another creates an excellent opportunity to clean and standardize descriptive metadata. The processes involved in moving large amounts of data from one platform to another lend themselves to automated analysis and remediation of metadata problems. The University of Houston (UH) Libraries established a Digital Asset Management System (DAMS) Implementation Task Force in early 2014 to explore possibilities for implementing a more robust repository architecture for the UH Digital Library. During the digital asset management system testing process, the UH Libraries Metadata Services Coordinator developed a set of scripts to programmatically access the data in the UH Digital Library through the existing digital asset management system API, create reports that were used to identify and correct problems, and lay the foundation for publishing UH Digital Library metadata as linked data. This project report discusses the background for the DAMS Implementation Task Force's work and the metadata quality improvements that resulted from it as part of a new Metadata Migration Project.

Keywords: metadata migration; quality control; digital asset management; automation; controlled vocabularies; linked data

1. Introduction

Metadata quality is an often overlooked or neglected aspect of digital repository development. In the excitement of setting up a repository infrastructure, the focus typically points to the software and hardware that allow institutions to publish digital collections on the World Wide Web, such as scanners, cameras, servers and turn-key content management system software. In the absence of trained metadata staff, descriptive metadata creation becomes a secondary activity that must be done in order to get a collection online rather than an essential process that facilitates effective discovery of a repository's resources.

Over time, as a repository's content grows, repository managers may realize that the quality of their descriptive data has suffered in the absence of careful attention to detail and consistent application of recognized standards. This is especially true when an institution explores opportunities for migrating data from one digital asset management system to another, as data analysis begins and decisions must be made regarding metadata transformations. This project report describes how the University of Houston (UH) Libraries leveraged the decision to test new digital asset management system software to analyze metadata in the UH Digital Library (UHDL), correct the problems it found, and prepare the UHDL descriptive metadata for publication as linked data.

2. Digital Asset Management System Evaluation

Since the launch of the UHDL in 2009, the UH Libraries have made thousands of rare and unique items available online using CONTENTdm, a proprietary digital asset management system owned and maintained by OCLC. While CONTENTdm helped the UH Libraries establish digital collections, the system has its limitations. The UH Libraries' digital initiatives have expanded, and the UHDL requires a more dynamic and flexible digital asset management system

that can manage larger amounts of materials in a variety of formats. The new digital repository infrastructure must also accommodate creative workflows and allow for the configuration of additional functionalities such as digital exhibits, data mining, cross-linking, geospatial visualization, and multi-media presentation. In addition, a system designed with linked data in mind will allow the UH Libraries to publish its digital collections as linked open data within the larger semantic web environment.

The *University of Houston Libraries Strategic Directions, 2013-2016* set forth a mandate to “work assiduously to expand our unique and comprehensive collections that support curricula and spotlight research. We will pursue seamless access and expand digital collections to increase national recognition” (p. 7). To fulfill the UH Libraries’ mission and the mandate of the strategic directions, a Digital Asset Management System (DAMS) Implementation Task Force was created to explore, evaluate, test, and recommend a more robust DAMS that can provide multiple levels of access to the UH Libraries unique collections at a larger scale. The collaborative task force consists of representatives from four library departments: Metadata & Digitization Services (MDS), Web Services, Digital Repository Services, and Special Collections.

3. Metadata Upgrade Project

Concurrent with the work of the DAMS Implementation Task Force, the Metadata Unit in MDS wrapped up a two year project to normalize and standardize the legacy descriptive metadata in the UHDL. The Metadata Upgrade Project was initiated in 2013 to systematically analyze the descriptive metadata in the UHDL, standardize Dublin Core field usage across the UHDL’s collections, and correct metadata content errors (Weidner et al., 2014). The analysis (Phase 1) and standardization (Phase 2) phases of the project produced a Metadata Dictionary (2014) input standard that guided the remediation work undertaken in the third phase, as well as metadata creation for new UHDL collections.

During the remediation phase (Phase 3) of the Metadata Upgrade Project, the Metadata Unit staff edited descriptive metadata for 54 collections comprising more than 9,100 digital objects. The Metadata Upgrade staff followed a workflow outlined at the beginning of the project. Tasks varied from collection to collection, depending on the state of the original metadata. Many tasks were accomplished through automation, such as aligning subject terms with controlled vocabularies (Weidner et al., 2014). After the Metadata Upgrade Project’s metadata remediation phase was complete, the Metadata Unit staff conducted an audit of the tasks outlined in the project plan. Anomalies were noted, along with tasks that fell outside of the original project scope, for a subsequent undertaking to further refine the descriptive metadata in the UHDL.

4. Systems Testing

In late 2014, the DAMS Implementation Task Force began testing two systems as part of its charge to select a new repository architecture for the UHDL: DSpace 4 and Fedora 3. Web Services installed both systems in a development environment, and test collections from the UHDL were selected for ingestion into both systems. Rather than start from scratch with the original files and spreadsheet metadata, the Metadata Services Coordinator developed a set of Ruby scripts that access the data in the UHDL through the CONTENTdm API. These “cdmeta” scripts harvest image, audio, and video files as well as descriptive data and transform the descriptive data into DSpace Dublin Core and Fedora FOXML metadata (Weidner, 2015). Using these scripts, metadata and files for the test collections were quickly produced in the ingest formats required by DSpace and Fedora.

Recognizing the potential for applying the same technique to the Metadata Upgrade Project’s authority control work, the Metadata Services Coordinator re-wrote the systems testing scripts as a Ruby library for object oriented access to the CONTENTdm API and created scripts that harvest names and subject terms in the UHDL. The “cdmeta_reports” scripts collate the harvested vocabulary data in plain text reports that list which objects are described by each term (Weidner,

2015). A second set of scripts filters the harvested lists of names and subject terms for unique values and writes those values to text files for each controlled vocabulary. Preliminary inspection of the vocabulary harvest files revealed common authority control problems, such as misspelled terms and multiple versions of the same name. Further inspection revealed terms that do not exist in the vocabulary to which they were assigned in the UHDL. Between the issues identified in the Metadata Upgrade Project audit and the controlled vocabulary terms harvest during systems testing, MDS recognized the need for a new project to prepare the UHDL's descriptive data for systems migration. As shown in Table 1, the work completed during the Metadata Upgrade Project and Systems Testing set the stage for the Metadata Migration Project that is currently underway.

TABLE 1. UH Libraries Metadata Projects Goals

Project	Goals
Metadata Upgrade	Standardize Metadata Schema Establish Input Standard Implement Controlled Vocabularies Correct Mistakes
Systems Testing	Develop Tools for Data Extraction Develop Tools for Analyzing Repository Data
Metadata Migration	Align Data with Controlled Vocabularies Prepare for Data Migration Prepare for Linked Data

5. Metadata Migration Project

The Metadata Migration Project at the UH Libraries began in early 2015 after the completion of the Metadata Upgrade Project. Expected to last until mid-2017, the project aims to build on the workflows and tools developed during the Metadata Upgrade Project and DAMS Implementation Task Force systems testing to further refine the UHDL's descriptive metadata in preparation for migration to a new digital repository architecture. The project will consist of iterative cycles of analysis and remediation to align controlled vocabulary terms with recognized authorities and prepare for linked data. After the new repository architecture has been implemented, the Metadata Migration Project will be complete, and all UHDL content will be migrated to the new system (Figure 1).

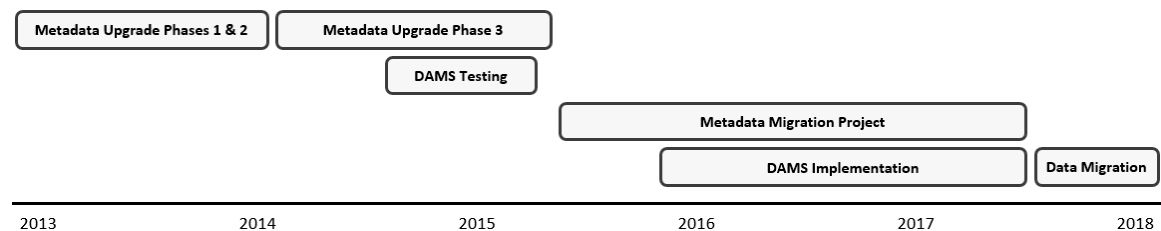


FIG. 1. UH Libraries Metadata Projects Timeline

5.1. Metadata Analysis

Using the cdmata reports scripts described in Section 4, the Metadata Services Coordinator compiled lists of all the subject terms and names in use in the UHDL and further separated all of the unique values into vocabulary specific lists. This began a stage of metadata analysis that required staff time and the development of additional tools to partially automate the verification of controlled vocabulary terms. Verification of subject terms and names followed a two-step

process designed to identify problems with the values in use in the UHDL and gather URIs for valid terms that will be used in future linked data applications. A different employee performed each step so as to guarantee the authoritative nature of the UHDL's confirmed authority links.

The first step's primary goal was to gather URIs from the source vocabulary for authorized terms in use in the UHDL. To accomplish this task quickly and accurately, the Metadata Services Coordinator wrote an AutoHotkey (2015) application that automated repetitive tasks and allowed Metadata Unit staff to focus on verifying content. The application parses a controlled vocabulary list and displays each unverified term in a dialog box (Figure 2). At the same time, the application opens a search for the term in the vocabulary's online user interface in a web browser. The user can position the dialog box in a convenient location on the screen so as to quickly verify whether or not the UHDL term matches the term in the source vocabulary. If a match is found, the user clicks the Yes button and the application instructs the user to navigate to the linked data web page for that term. In the case of the Art and Architecture Thesaurus (AAT), that page is the Semantic View for Getty's Linked Open Data Vocabularies (Getty Vocabularies, 2015), as shown in Figure 3.



FIG. 2. Authority Verification Application Dialog Box

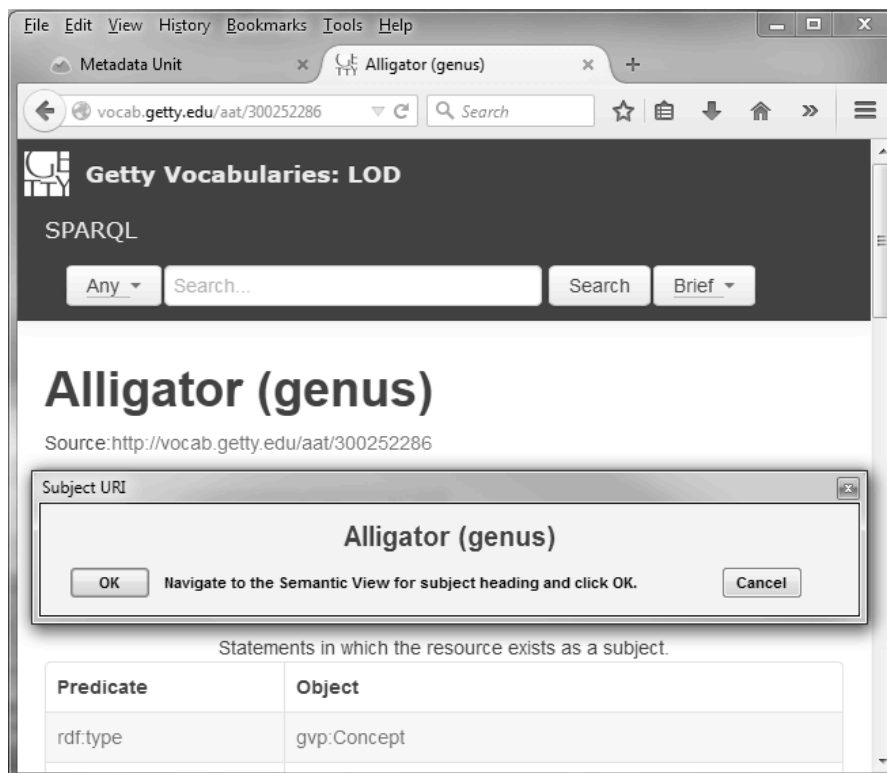


FIG. 3. Authority Verification Application & AAT Semantic View

After the user clicks OK, the application automatically harvests the subject heading URI, closes the tab in the web browser, and begins the process again for the next unverified term.

Verified terms and their associated URIs are recorded in a tab delimited text file. If the user discovers a problem with the term in use in the UHDL, clicking No in the initial dialog opens a second dialog, shown in Figure 4, which provides radio button options for indicating what is wrong. Common problems include misspelled headings and headings that have less or more information than the authorized form. Problem terms are recorded in a separate text file for further analysis and remediation work described in the next section. The second step in the controlled vocabulary term verification process utilizes a similar AutoHotkey application that displays a term in a dialog box, opens the linked data web page for that term, and asks the user to verify that the term in the dialog box matches the term on the web page. Any problems discovered during this stage are recorded in a separate text file, and the twice-verified tab delimited list of terms and their associated URIs are ready to be reformatted for use as linked data.

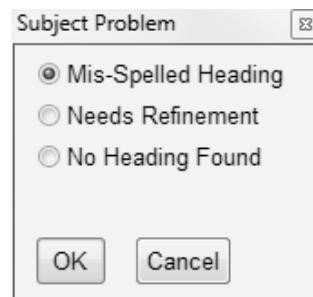


FIG. 4. Authority Verification Application Problems Dialog

5.2. Metadata Remediation

Despite the best efforts of the Metadata Upgrade Project, the programmatic harvest and analysis of the controlled vocabulary terms in use in the UHDL revealed many problems remaining to be corrected. The problems ranged in difficulty from misspelled subject headings to headings assigned out of context. Some of the context problems occurred because of an automation application used during the Metadata Upgrade Project that only allowed for one mapping from an alternate subject vocabulary to LCSH (Weidner et al., 2014). Other cases were the result of inadequate training of staff in descriptive practice and lack of effective metadata quality control at various times since the creation of the UHDL in 2009.

In order to make the large and complex remediation process more manageable, the Metadata Services Coordinator wrote an AutoHotkey script that cross-references the list of problems compiled during the authority verification process with the list of all subject terms in use in the UHDL. The script creates a new tab delimited text file for each controlled vocabulary that lists the subject term error and URLs to each object in the UHDL with that term in its metadata record, as shown in Figure 5. When viewed in Notepad++ (2015), the URLs became clickable links that Metadata Unit staff used to quickly locate the objects that required attention. The Metadata Unit Staff added authorized terms and URIs to the same tab delimited file after correcting the errors within the CONTENTdm Project Client.

	UHDL Term	Error	Authorized Term	URI	UHDL Object Links
250	0 banks	Needs Refinement			
251	1 banks (financial institutions)			http://vocab.getty.edu/aat/300005214	
252	historic_houston	aat banks		http://digital.lib.uh.edu/collection/p15195coll12/item/161	
253	historic_postcards	aat banks		http://digital.lib.uh.edu/collection/p15195coll116/item/358/show/348	
254	houston_magnolia_city	aat banks		http://digital.lib.uh.edu/collection/p15195coll11/item/260	
255	houston_magnolia_city	aat banks		http://digital.lib.uh.edu/collection/p15195coll11/item/282	
256	houston_magnolia_city	aat banks		http://digital.lib.uh.edu/collection/p15195coll11/item/329	

FIG. 5. Subject Term Errors with Authorized Forms and Links to UHDL Objects

For name authority reconciliation, the Metadata Unit leveraged a set of open source OpenRefine scripts that automatically harvest URIs from the Library of Congress Name Authority File (LCNAF) by querying the Virtual International Authority File (Carruthers, 2015). After separating the UHDL name lists into personal and corporate names, the OpenRefine scripts produced lists of matches with LCNAF URIs. The Metadata Services Coordinator developed an AutoHotkey script to divide all of the names into three categories: probable matches, questionable matches, and unmatched terms. Of the 1,223 unique names in the UHDL's descriptive metadata, the OpenRefine scripts found probable matches for 355 names, questionable matches for 347, and 521 names remained unmatched. Similar to the verification and remediation work for the UHDL's subject terms, AutoHotkey apps were developed to confirm linked data URIs and identify records that required metadata corrections in the name fields.

5.3. Linked Data and DAMS Implementation

An integral part of the Metadata Migration Project is preparing for the linked data environment. As previously mentioned, the Metadata Unit staff used a variety of applications to systematically harvest and verify URIs for authorized subject and name terms in a number of controlled vocabularies. Whenever possible, the process of recording the URI was automated to avoid copy and paste errors. This was accomplished with AutoHotkey by copying the text in the browser's address bar, as shown in the AutoHotkey function in Figure 6.

```

124 GetURI:
125     WinGetPos, xSubURI, ySubURI,,, Subject URI
126     winX := xSubURI
127     winY := ySubURI
128     Gui, 2:Destroy
129
130     WinActivate, LC Linked Data Service
131     Sleep, 50
132     Send, ^l
133     Sleep, 50
134     Send, ^c
135     Sleep, 50
136     StringTrimRight, uri, Clipboard, 5
137     IfNotInString, uri, vocabulary/graphicMaterials
138     {
139         InvalidURI = 1
140     }
141 Return

```

FIG. 6. AutoHotkey Function to Harvest URI from Web Browser Address Bar

Eventually these links will enter the UHDL metadata to assert a relationship between the object and a subject term maintained in an external vocabulary. MDS is currently investigating the deployment of a vocabulary server to facilitate the consistent use of controlled vocabulary terms in the UHDL and throughout the UH Libraries (TemaTres, 2015). The UH Libraries will soon be implementing a new DAMS infrastructure based on Fedora 4 (2015), which conforms to the W3C recommendation for Linked Data Platforms (2015). Because of the work accomplished during the Metadata Migration Project, the UH Libraries will be in a good position to quickly publish our digital objects with links to external vocabularies when the migration to Fedora occurs.

6. Conclusion

The UH Libraries Metadata Migration Project is a natural continuation of the Metadata Upgrade Project. The improved quality of metadata, with URIs for controlled vocabulary terms, will prepare the UH Libraries for a smooth data migration to a new digital asset management system designed for the linked data environment. The implementation of the new system based on Fedora 4 will allow us to publish our digital collections as linked open data and open up new possibilities for effective use and re-use of the UH Libraries unique digital collections.

References

- AutoHotkey. (2015). Retrieved April 7, 2015, from <http://www.autohotkey.com>.
- Carruthers, Matt. (2015). LCNAF-Named-Entity-Reconciliation GitHub repository. Retrieved July 14, 2015, from <https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation>.
- Fedora Repository. (2015). Retrieved April 7, 2015, from <http://fedorarepository.org>.
- Getty Vocabularies: Linked Open Data. (2015). Retrieved April 6, 2015, from <http://vocab.getty.edu>.
- Metadata Dictionary. (2014). University of Houston Digital Library. Retrieved March 31, 2015, from <http://digital.lib.uh.edu/about/metadata>.
- Notepad++. (2015). Retrieved April 7, 2015, from <http://notepad-plus-plus.org>.
- TemaTres. (2015). Retrieved July 21, 2015, from <http://www.vocabularyserver.com>.
- University of Houston Libraries Strategic Directions, 2013-2016. (2013). Retrieved March 31, 2015, from <http://info.lib.uh.edu/sites/default/files/docs/strategic-directions/2013-2016-libraries-strategic-directions-final.pdf>.
- W3C. (2015). Linked Data Platform 1.0 Recommendation. Retrieved April 7, 2015, from <http://www.w3.org/TR/ldp>.
- Weidner, Andrew, Annie Wu, and Santi Thompson. (2014). Automated Enhancement of Controlled Vocabularies: Upgrading Legacy Metadata in CONTENTdm. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2014, 167-172.
- Weidner, Andrew. (2015). cdmata GitHub repository. Retrieved April 1, 2015, from <https://github.com/metaweidner/cdmata>.
- Weidner, Andrew. (2015). cdmata_reports GitHub repository. Retrieved April 1, 2015, from https://github.com/metaweidner/cdmata_reports.

Understanding Metadata Needs when Migrating DAMS

Ayla Stein
University of Illinois at
Urbana-Champaign, USA
astein@illinois.edu

Santi Thompson
University of Houston,
USA
sathompson3@uh.edu

Abstract

This study identifies and explores metadata needs associated with migrating to a new Digital Asset Management System (DAMS). Drawing upon results from a 2014 survey, titled “Identifying Motivations for DAMS Migration: A Survey,” this paper analyzes survey questions related to metadata, interoperability, and digital preservation. Results indicate three distinct metadata needs for future system development, including support for multiple or all metadata schema, metadata reuse, and digital object identifiers. While some of these needs resemble long-standing conversations in the professional literature, others offer new areas for system development moving forward.

Keywords: metadata, digital asset management systems

1. Introduction

In the last two decades, digital asset management systems (DAMS) have become important tools for collecting, preserving, disseminating, and making discoverable digitized and born digital content to library users. During that time libraries have selected a variety of DAMS to manage their digital assets, including proprietary systems (Ex Libris’ DigiTool and OCLC’s CONTENTdm), open source platforms (Greenstone, Fedora, Islandora, and DSpace), and homegrown solutions. Over time libraries have begun re-assessing DAMS based on the changing needs of users, the expanding skill sets of librarians and staff, and the evolution of web technologies. As libraries engage in this process, some choose to migrate from one DAMS to another.

The data referenced in this paper is drawn from “Identifying Motivations for DAMS Migration: A Survey,” which identified thirteen topical categories for migrating from one digital asset management system (DAMS) to another. Researchers focused the survey on systems used to provide access to primary source research materials. The scope emphasized that the survey did not focus on systems used *exclusively* as institutional repositories, which the researchers define as repositories that provide access to university scholarship. This paper analyzes a subset of the responses which focus on the topics Metadata Standards, Interoperability, and Preservation. In the survey the researchers defined each of the three categories as:

- Metadata Standards: The “New DAMS’s” support of established metadata standards, user generated metadata, and linked data technologies.
- Interoperability: The “New DAMS’s” ability to export metadata into other DAMS and digital program environments. The “New DAMS” should support international and/or industry standards for interoperability, including OAI-PMH, Z39.50, and SRU/SRW protocols.
- Preservation: The integration of preservation strategies into the “New DAMS”, including fixity verification and the creation of checksum values, backups, synchronization, and/or the generation of archival information packages (AIPs).

The researchers believe that results from this data may give insight to the question, “What are the metadata needs for migrating from one DAMS to another?” Understanding these needs could help align future DAMS development and adoption with emerging metadata trends and initiatives.

2. Literature Review

Metadata is a core element to any library DAMS. This literature review examines works that focus on the relationship between metadata and DAMS functionality in order to compare established practices, identified gaps in the literature, and emerging needs from survey results.

Attention to relationship between metadata and library DAMS has been diverse. Some information professionals have addressed broad ways that metadata supports core DAMS functionality. Payette (1998) identified several library functions, including resource discovery, access and use, preservation and administration, and persistent identifiers. Others have focused on specific tools and features. Lagoze et. al. (2005) discussed how metadata automation comes in several different flavors, including: detection of embedded metadata within ingested digital objects and auto-generated metadata values. Tools to create and manage both traditional and non-MARC metadata are another significant concern. In Zeng et al.'s 2009 survey, they found that survey respondents were concerned with a lack of metadata tools that are easy to use and do not require a steep learning curve.

User contributions is another metadata feature in library DAMS that has received attention in the professional literature. In order for DAMS to meet the needs of users, they need to upgrade to Web 2.0., which is characterized by user contributions and interactions with online content (Beal, n.d.). In their comparison survey of DAMS, Andro et. al. (2012) identified several systems that enabled users to make contributions to metadata, either through the process of “annotating” or “commenting” (p. 82). Others have focused on issues that arise from implementing user contributions in a DAMS. Lagoze et al. (2005) discuss how user contributed content creates complications for system distinctions between metadata and data. They note “...one of the useful forms of contextual information is annotations. Are these metadata (about something) or data in their own right? There is no one answer, but an architecture that imprints the distinction between data and metadata makes it difficult to deal with such ambiguities” (p. 6). Still other parts of the literature emphasize curatorial and user engagement possibilities. For example, crowd-sourced additions might augment or even replace time-intensive and expensive metadata creation and maintenance work (Mitchell and Gilbertson 2008).

Some studies have identified intersections between metadata and DAMS development that need further research and support. In their comparison of 10 DAMS, Andro et.al (2012) compared how systems supported multiple metadata schema, including non-traditional library schema. The authors discovered that all or most systems supported some degree of “library” metadata (including DC, MODS) and “archives” (EAD) (p. 80). However, less than half of the systems supported “research,” “learning,” and “photo” metadata (p. 80). This research suggests that current systems lack support around metadata schema and functions that describe research and other activities outside of the library environment. Additionally, Goh et. al. (2006) proposed that systems should support multiple metadata schema since virtually all of the systems evaluated in their study supported only core standards (such as MARC21 and Dublin Core) (p. 367). Furthermore, Han et. al (2010) suggested that future research could better configure complex objects in CONTENTdm to maximize discoverability and interoperability (p. 77).

Finally, metadata contributes to DAMS functionality through enabling interoperability. Because DC is a flexible, simple schema, it is well suited for promoting interoperability among other systems. Han et. al. (2010) argued that drawing on best practices could help promote interoperable metadata as well as eliminate metadata problems derived from inconsistencies in

localized practice (p. 74-75). Zeng et. al. (2009) noted that there is also significant interest in designing systems that can natively handle or map between different metadata schemas. Additionally, Lagoze et. al. (2005) wrote that “we should be wary of throwing out collections of cataloging records, and ignoring the value that uniform metadata has for ‘order making’ over heterogeneous information. However, we need to incorporate these catalog records into a richer foundation that represents...complex relationships and a host of other complexities” (p. 6).

Linked data technology is widely considered to be the solution to non-metadata centric systems (Solodovnik 2011). While interoperability issues may be somewhat ameliorated by the implementation of linked data, as it currently stands, the legacy methods of developing metadata vocabularies, in disciplinary silos, is being carried over to the Semantic Web: “a major source of interoperability problems on the Semantic Web is still due to the use of different value vocabularies supporting metadata descriptions in different linguistic communities” (Solodovnik 2011, p.10). It’s clear that the use of Linked Data technologies in and of themselves will not be enough to promote interoperability. It will require cross-disciplinary and inter-institutional collaboration. The success of the schema.org vocabulary could arguably be attributed to the fact that it was developed and implemented by the three largest search engine corporations: Google, Yahoo!, and Bing (O’Connor 2011).

The results of this survey build upon many of these themes, including the continued need for supporting multiple metadata schema, sharing data among systems in new ways, and the future role of linked data. It also begins to expand the discussion around differing development areas, including metadata reuse among users and the role of digital object identifiers in library DAMS intended to curate and make accessible digitized special collections materials.

3. Methodology

To complete this study, researchers analyzed a subset of data from a larger investigation that seeks to identify motivations for migrating from one DAMS to another. Using a survey as their instrument, they solicited responses by emailing calls for survey participation to eight listservs related to digital curation from July through September 2014.¹ In order to qualify for the survey, respondents had to fulfill one of the following three eligibility categories:

1. Institutions had completed migration from the “Old DAMS” to the “New DAMS”
2. Institutions were currently migrating from the “Old DAMS” to the “New DAMS”
3. Institutions selected a “New DAMS” but had not started the migration process
- 4.

If institutions selected “none of the above,” the software automatically ended the survey. Since the researchers solicited anonymous responses from listserv subscribers, they did not have the information needed to calculate a response rate. Once initiated, the survey had a completion rate of 47%. After removing ineligible entries, the researchers had 49 responses to analyze for this study. Over half of the eligible responses came from academic libraries. For more information, see Table 1: Which of the following best describes your library?

¹The listservs included: The Code4Lib main listserv; DigLib, the International Federation of Library Associations (IFLA)'s digital library focused listserv; DigiPes, an American Library Association (ALA) listserv focused on digital preservation issues; Archives and Archivists, the main listserv for the Society of American Archivists; the Research Data Access and Preservation (Rdap) focused listserv from the Association of Information Science and Technology (ASIS&T); DLF-ANNOUNCE, a listserv from the Digital Library Federation; pasig-discuss, the discussion listserv for ASIS&T's Preservation and Archiving Special Interest Group (PASIG); and acr-igdc-l, the listserv for the Association of College and Research Libraries' (ACRL) Digital Curation Interest Group.

TABLE 1: Which of the following best describes your library?

Response Type	Total Number of Responses	%
Academic Library	30	61
Research Library	8	16
Public Library	4	8
Special Library	2	4
Special Collections Libraries or Archives	2	4
Government Library	2	4
Other	1	0
Museum Library	0	0

To create the survey, the researchers crafted specific questions around thirteen topics related to DAMS evaluation, including:

- Implementation & Day-to-Day Costs
- User Administration
- Organizational Viability
- Technical Support
- System Administration
- Extensibility
- Information Retrieval & Access
- Content Management
- Preservation
- User Interface Customization
- Interoperability
- Reputation
- Metadata Standards

Survey questions for these topics were designed to be either a Likert scale of 1 [Not Important] to 4 [Very Important] or select all that apply. The survey asked for key demographic information to help the researchers understand how institutions prioritized potential motivations. Demographic questions required respondents to select and/or self-identify the “Old DAMS” and the “New DAMS.” Next, the survey asked respondents to choose the top five motivations from one of the thirteen topics and then prioritize those five selections in order of importance. At that point, respondents answered questions from the five topics they identified.

Since the scope of this paper is to understand the relationship between metadata needs and DAMS migration, the researchers identified questions that addressed metadata features and functionality. Researchers used the survey reports feature in Qualtrics to generate descriptive statistics for the selected questions, including total amount, statistical mean, and standard deviation. They drew upon these reports to formulate conclusions and identify future research areas.²

² The Qualtrics reports also included minimum and maximum values, as well as variance.

4. Results

Analyzing the data³, researchers determined whether certain metadata features were important or not important to respondents.

TABLE 2: Survey Questions Related to DAMS Metadata Features and Functionality.

Question	Total Number of Responses	Mean	SD
The ability to allow other digital library environments to harvest its content	16	3.75	0.45
The ability to support multiple metadata schema	22	3.68	0.57
The "New DAMS" has the ability to export all or part of the metadata for reuse	16	3.50	0.82
The ability to support local metadata standards and practices	22	3.32	0.95
The new dams supports digital object identifiers	22	3.23	0.97
The new dams supports linked data technologies	22	2.82	1.10
The ability to support user generated metadata such as tags or folksonomies	22	2.59	1.05
The new dams automates metadata creation	10	2.50	1.18
The new dams supports personal digital identifiers	21	2.24	0.94

As can be seen in Table 2, researchers considered results that registered mean responses higher than 3.0 and a standard deviation of less than 1.0 to be important considerations for institutions migrating to a new DAMS. These included:

- "The ability to support multiple metadata schema"
- "The ability to support local metadata standards and practices"
- "The new DAMS supports digital object identifiers"
- "The ability to export all or part of the metadata for reuse"
- "The ability to allow other digital library environments to harvest its content"

Alternatively, researchers considered results that registered mean responses lower than 3.0 and/or a standard deviation at or above 1.0 to be less important considerations for institutions migrating to a new DAMS. These included:

- "The new DAMS supports linked data technologies"
- "The ability to support user generated metadata such as tags or folksonomies"
- "The new DAMS automates metadata creation"
- "The new DAMS supports personal digital identifiers"

Other responses demonstrate the diverse needs that future DAMS should address to remain relevant to the cultural heritage community.

TABLE 3: Detailed Survey Questions Related to DAMS Metadata Features and Functionality

Survey Question	Survey Answer	Total Number of Responses	%
What descriptive metadata standards/schema did you desire the "New	Dublin Core	19	90
	MODS	16	76
	EAD	12	57
	MARC	10	48

³ Researchers are actively working with the data from this survey to complete another manuscript for publication. However, data are available upon request to the authors. Once published, the researchers will make the data from this project freely accessible via a repository.

DAMS" to support?	VRA Core	7	33
	PB Core	3	14
	DDI	3	14
	All Schema/Schema-less	3	14
	GNS	1	5
TOTAL RESPONSES		74	
What metadata did you desire the "New DAMS" to automatically create?	Technical metadata	8	100
	Preservation metadata	5	63
TOTAL RESPONSES		13	
What administrative, preservation, structural, and/or technical metadata standards did you desire the "New DAMS" to support?	METS	18	90
	PREMIS	15	75
	TEI	8	40
	VRA Core	5	25
	MIX	2	10
	PB Core	2	10
TOTAL RESPONSES		50	
What interoperability methods and/or standards did you desire the "New DAMS" to support?	OAI-PMH	14	88
	APIs	9	56
	Z39.50	6	38
	SRU/SRW	3	19
	OAI-ORE	1	6
	SPARQL	1	6
TOTAL RESPONSES		34	
What linked data technologies did you desire the "New DAMS" to support?	RDF/XML	16	89
	JSON	10	56
	Rich Snippets/Rich Data	2	11
	Other	1	6
TOTAL RESPONSES		29	
What digital object identifiers did you want the "New DAMS" to support?	doi	17	60
	ezid	4	14
	ARK	3	11
	handle	2	7
	urn:nbn	1	4
	local	1	4
TOTAL RESPONSES		28	
What personal digital identifiers did you want the "New DAMS" to support?	ORCID	12	46
	ARK	5	19
	ResearcherID	4	15
	Other	3	12
	MADS Authorities	1	4
	ISNI	1	4
TOTAL RESPONSES		26	

While Dublin Core was the most popular response for descriptive metadata, several other standards/schema also had a high number of responses, which suggests that future systems should support multiple descriptive schemas. Additionally, the researchers received several free text responses that said DAMS should support all metadata schemas or should be schema-less. All respondents desired technical metadata to be automatically created by the DAMS. A majority of participants also expected that preservation metadata would be collected systematically. Future systems should support and generate METS records, as well as document PREMIS events as part of their core functionality. For interoperability, respondents favored using OAI-PMH and APIs over other methods to share metadata with other systems. In regards to linked data, RDF/XML and JSON are the most popular serialization formats for expressing metadata as linked data. Concerning identifiers, DOIs appear to be the most widely needed object identifiers for future systems. Additionally, if systems choose to support personal digital identifiers (PDIs), they

should particularly consider ORCID, as well as authority identifiers such as ISNI, and authority schemas like the Metadata Authority Description Schema (MADS).

5. Discussion

Researchers drew upon response data from several survey questions to answer the research question: “What are the metadata needs for migrating from one DAMS to another?” Creating systems that support all or multiple types of metadata schema was one important need derived from survey results. The responses to the survey question “The ability to support multiple metadata schema” showed respondents desired more metadata flexibility from DAMS. The follow up questions “What descriptive metadata standards/schema did you desire the “New DAMS” to support?” and “What administrative, preservation, structural, and/or technical metadata standards did you desire the “New DAMS” to support?” affirm metadata practices that are commonly used today among institutions. For example, Dublin Core, METS, and PREMIS remain the most popular schema overall for a New DAMS to support. Since the survey did not ask respondents to explain their preferences, researchers could only speculate as to why those completing the survey selected these specific schemas. Survey results also showed that a majority of respondents desired support for other metadata schema. In addition to desiring support for DC, respondents also favored MODS, and EAD for descriptive metadata, while a still sizable number also preferred MARC and VRA Core.⁴ Combining these results with the favorable support of another survey question, “The ability to support local metadata standards and practices,” suggests a need for future systems to support multiple or all schema (either locally-derived or based on formal standards) as Goh et. al. (2010) argued (p. 367).

Another need that emerged from the survey results focused on facilitating library metadata reuse by both systems and users. The responses to the survey question related to “The ability to allow other digital library environments to harvest its content” suggested that respondents still highly valued the ability to make their data interoperable with other library DAMS. A follow up question, “What interoperability methods and/or standards did you desire the “New DAMS” to support?” showed that OAI-PMH remains the most popular aggregation method for respondents, surprising researchers who thought the growing system development around APIs would have made it the most popular method. Despite libraries growing comfort in the technological realm, implementing new technologies such as APIs still requires specialized knowledge and skills, which may be why established protocols such as OAI-PMH are still in high demand. There may also be a desire to support technologies developed within the library domain and some inherent resistance to external innovations. With limited resources and time, librarians may prefer to stay with the technologies they have helped to create and support over time.

Complementing system reuse, results from the survey question “The “New DAMS” has the ability to export all or part of the metadata for reuse” showed how respondents favored system functionality around user reuse. Often metadata records contain rich contextual information about digital objects that, in itself, can be valuable data for research. Because the amount of attention focused on reusing data, from data sets to metadata in digital humanities projects, has increased over the last several years, the researchers were not surprised by this need. Since most of the literature dedicated to selecting DAMS and to the role that metadata plays in DAMS functionality do not address user reuse of metadata, the researchers believe that a gap exists in the literature around designing DAMS for metadata reuse by the user; this a gap should be addressed in future research.

A third need focused on future DAMS supporting digital object identifiers. Results from a survey question that explored “The new dams supports digital object identifiers” suggested that

⁴ Schema focused on particular formats or content types (VRA Core and PB Core, for example) were not as highly selected; if it is not possible for system to support all schema, it is unclear just how integrated future systems should be with these schema.

respondents desired a future system that has the capability to generate identifiers for digital objects. A follow up question, “What digital object identifiers did you want the “New DAMS” to support?” showed that respondents favored Digital Object Identifier System identifiers (DOIs) specifically, which surprised the researchers because of the cost implications related to DOIs, as well as the lack of anecdotal evidence of libraries adopting DOIs for their digitized collections. Readers should note that there are some limitations around the results of this particular question based on an error in the survey instrument. Researchers included ezid as possible response for the follow up question related to digital object identifiers. Since ezid mints identifiers (dois and ARKs), it should not appear in the question. Additionally, the scope of the survey was based on DAMS intended to curate digitized special collections content. However, some institutions may have one unified DAMS that fulfills multiple purposes, including disciplinary or institutionally-based repositories, which have a wider adoption of digital object identifiers. In any case, future research on the role of digital object identifiers in digital library/digital collections environments should be explored further.

While three needs emerged from the survey data, the researchers concluded that the response data to the other topics related to metadata and future DAMS development could not be applied to the research question because these are areas that require more in-depth research and investigation.

The responses to the survey question “The New DAMS supports linked data technologies” indicated a lack of consensus on whether or not linked data technologies were considered necessary for New DAMS. While fifteen respondents indicated that support of linked data technologies were considered ‘important or very important’, seven respondents indicated ‘not important’ or ‘somewhat important’. The lack of consensus reflects the present status of applied linked data technologies. Until relatively recently, linked data was, and still often is, an abstract or intangible concept. While research, investigation, and infrastructure development on library linked data has been underway for several years (Baker et. al. 2005; Library of Congress, n.d.), it was not until the release of Fedora 4 (DURASPACE 2014), and to a lesser extent Kuali OLE (Kuali n.d.), that native linked data library systems became readily available. Even between these two systems, only Fedora 4 could function as a DAMS. There is still a significant amount of work that needs to be accomplished before linked data technology is within reach of most libraries.

The responses to the survey question “The ability to support user-created metadata such as tags or folksonomies” also indicated a lack of consensus. Responses were almost evenly distributed, with ten respondents indicating it was ‘Not important or somewhat important’, and twelve indicating ‘important or very important. These results were somewhat surprising in light of the significant interest and optimism regarding user-created tags in the literature (Lagoze et. al. 2005; Mitchell and Gilbertson 2008). The formation of questions may have also impacted results. The researchers focused entirely on user-created vocabularies, and did not include examples of added-value metadata, e.g. annotations. The researchers suspect that the type of user-created metadata needed in DAMS has changed over time, (especially with the proliferation of tablets, “phablets” (Oxford English Dictionary 2015), and touchscreens) and research-oriented user-metadata features, like highlighting and annotating, would be rated more highly. This topic is an area of future investigation that the researchers hope to explore further with institutional, data, and scholarly repositories.

Responses to the question “The New DAMS automates metadata creation” indicate that participants do not consider automated metadata creation to be a required function of the New DAMS. These results were surprising to the researchers given the attention that the literature paid to the varieties of metadata automation (Lagoze et. al. 2005). The researchers believe that the results are partially due to poor wording of the question, which does not reflect multiple types of metadata automation. A follow-up question, “What metadata did you desire the “New DAMS” to automatically create?” asked respondents to select-all-that-apply with possible responses of ‘technical metadata’, ‘preservation metadata’, or ‘other/free text’. This question did not clarify

what researchers meant by automated metadata creation. A more appropriate question to ask would have focused on specific use cases for automated metadata creation.

Responses to the question “The New DAMS supports personal digital identifiers” conclusively indicate that personal digital identifiers (PDIs) are not necessary for New DAMS to support.⁵ This result did not surprise the researchers given that PDIs such as ORCID and ResearcherID are far more prevalent in institutional and scholarly DAMS than those focused on digital library collections. The connection between DAMS and PDIs is an area of future inquiry for the researchers.

Researchers have identified several limitations with the composition of the survey and the results derived from it. Because there is no definitive DAMS registry encompassing all libraries, the researchers cannot determine whether or not the results are statistically significant. Furthermore, the data are not necessarily based on a representative or random sample. Since researchers relied on voluntary participation from those who subscribed to certain listservs, they have no way of knowing the total number of possible participants or calculating a response rate.

6. Conclusion

The purpose of this investigation was to understand metadata needs when migrating from one DAMS to another. After analyzing both the existing literature and the survey results, the researchers have identified three specific needs:

1. Support for multiple or all metadata schema
2. Support for metadata reuse among other library DAMS as well as among users
3. Support for digital object identifiers

Viewed as metadata use cases for future DAMS developers, including both open source and proprietary, these three needs indicate that future DAMS should continue to embrace flexibility in metadata creation, management, export, and interoperability. In some ways, they mirror long-standing conversations in the professional literature. The desire to accommodate multiple schema and share it with a variety of library systems are not new or under-researched areas within the library profession; , however, these results do suggest that librarians and system developers have yet to bridge critical functionality gaps. To address these needs, conversations around metadata should be occurring from the earliest stages of system planning and development. Likewise, metadata specialists should be involved at all stages, from design to migration. Combining these needs with the desire to expand metadata reuse for users and to generate DOIs for rare and unique digitized materials offers a variety of development areas moving forward.

References

- Andro, Mathieu, Emmanuelle Asselin, Marc Maisonneuve (2012). Digital libraries: Comparison of 10 software. Library Collections, Acquisitions, & Technical Services 36.
- Baker, T., E. Bermes, and A. Isaac (2005). W3C Library Linked Data Incubator Group. Retrieved April 9, 2015, from <http://www.w3.org/2005/Incubator/lld/>
- Beal, V. (n.d.). What is Web 2.0? Webopedia. Retrieved April 5, 2015, from http://www.webopedia.com/TERM/W/Web_2_point_0.html.
- DURASPACE (2014). Now available: Fedora 4 production release - not your dad's fedora. Retrieved April 10, 2015, from <http://duraspace.org/articles/2394>.
- Goh, Dion Hoe-Lian, Alton Chua, Davina Anqi Khoo, Emily Boon-Hui Khoo, Eric Bok-Tong Mak and Maple Wen-Min Ng (2006). A checklist for evaluating open source digital library software. Online Information Review 30, no. 4.

⁵ However, several free text responses indicated the desire for systems to support authority files, including universal authority files, such as ISNI, and locally developed authorities through the support of MADS.

- Han, Myung-Ja, Sheila Bair, and Jason Lee (2010). Creating metadata best practices for CONTENTdm users. Proceedings of the International Conference on Dublin Core and Metadata Applications.
- Kuali (n.d.). Describe and manage module. Retrieved April 10, 2015, from <http://www.kuali.org/ole/modules/describe-manage-entity>.
- Lagoze, Carl, Dean Krafft, Sandy Payette, and Susan Jesuroga. (2005, November). What is a digital library anyway, anymore? Beyond search and access in the NSDL. D-Lib Magazine, 11 (11). Retrieved, January 10, 2007, from <http://www.dlib.org/dlib/november05/lagoze/11lagoze.html>.
- Library of Congress (n.d.) BIBFRAME - Bibliographic Framework Initiative. Retrieved April 10, 2015, from <http://www.loc.gov/bibframe/>.
- Mitchell, E., and K. Gilbertson, (2008). Using open source social software as digital library interface. D-Lib Magazine, 14, 1–11. Retrieved from <http://www.dlib.org/dlib/march08/mitchell/03mitchell.html>.
- O'Connor, M. (2011, July 20). schema blog [Blog]. Retrieved March 19, 2015, from <http://blog.schema.org/2011/07/on-june-2-nd-we-announced-collaboration.html>.
- Oxford English Dictionary (2015) Phablet. Retrieved April 10, 2015, from <http://www.oxforddictionaries.com/definition/english/phablet>.
- Payette, Sandra (1998 October 21). Metadata for digital libraries: a functional approach. Cornell Digital Imaging Workshop.
- Solodovnik, I. (2011). Metadata issues in Digital Libraries: key concepts and perspectives. *JLIS.it* 2 (2) (2011, December): 1-27. DOI: 10.4403/jlis.it-4663.
- Zeng, M. L., Lee, J., & Hayes, A. F. (2009). Metadata decisions for digital libraries: a survey report. *Journal of Library Metadata* 9 (April 2013), 173–193. doi:10.1080/19386380903405074.



Current Developments in Metadata for Research Data— Session 4

Dublin Core Usage for Describing Documents in Brazilian Government Digital Libraries

Diego José Macêdo
Instituto Brasileiro de
Informação em Ciência e
Tecnologia, Brasil
diegomacedo@ibict.br

Milton Shintaku
Instituto Brasileiro de
Informação em Ciência e
Tecnologia, Brasil
shintaku@ibict.br

Ronnie Fagundes de Brito
Instituto Brasileiro de
Informação em Ciência e
Tecnologia, Brasil
ronniebrito@ibict.br

Abstract

Digital libraries are increasingly common, being developed by government agencies to disseminate and preserve the documentation produced by its employees. This proposes a challenge in describing this type of documents, dealing official aspects in tools that are originally designed for bibliographic and scientific documents. In this sense, our objective is to verify how digital libraries, linked to the executive, legislative and judiciary Brazilian powers, are describing its documents collections. A study with descriptive and qualitative characteristics reveals the great adoption of DSpace software for creating these digital libraries and Dublin Core to describe the documents, showing DSpace and metadata schema adaptability for nonacademic document types. Thus, one contributes to the discussion on the use of Dublin Core to describe various types of documents on the Internet.

Keywords: Government digital library; Dublin Core; Government Agency.

1. Government Digital Libraries

With the change of the physical medium on paper for publication in electronic format, digital libraries have become the locus for preservation and access to documentation of an institution. With this, Brazilian governmental institutions created digital libraries in order to provide transparency to their activities, providing access to the full content of its documentation, creating a scenario where institutions use tools originally designed for the dissemination of scientific information in the dissemination of governmental information.

Many institutions have been using tools developed in free software, especially DSpace much by the support of the Brazilian Institute of Information Science and Technology (IBICT), which disseminates and supports this tool. This is also due to the government policy for free software adoption, which significantly changed the business of IT sectors, where development has been gradually replaced by adjustment of free tools.

This approach saves time and resources, since there is a large supply of free tools, with the most varied purposes. Some tools have a specific purpose and are being used for other purposes, such as DSpace, originally designed for academic repositories and used in other scenarios.

Another point to collaborate with the dissemination of government documents it related to the fulfillment of requisites defined by Law No. 527 of 18 November 2011, in which the agencies linked to the Brazilian government must make non-sensitive documents freely available. This law guarantees Brazilian population unrestricted access to governmental documents, regardless of support, encouraging the use of tools that support the digital distribution of documents, such as digital libraries.

Digital libraries are a dedicated tool for dissemination of scientific and technological documentation and have flow and structure aimed at managing these documents, which has well established forms of classification and cataloging. A challenge is posed to librarians, archivists and documentation developers in the description of processes of governmental documents in digital libraries.

In the bibliographical studies, there are few studies regarding government documentation, to the extent that many researchers classify them as archival documents. However, manuals, technical reports and other documents of institutional memory have bibliographic aspects, but not always receive adequate treatment in government agencies. So, these documents are not always disseminated, even with relevant information that could be reused.

In this context, the present study aims to analyze the use of Dublin Core metadata schema in the description of documents in digital libraries developed with DSpace and linked to Brazilian government agencies belonging to the executive, legislative and judicial branches. It analyzes their document's metadata and describes the strategies used for the representation of their collections.

2. Methodology

The study has descriptive characteristics, that is aimed to characterize populations or phenomena and suitable to describe scenarios (Gil, 2006). In line with the objective of analyzing the use of Dublin Core in the description of government documents, the research provides a survey of the Brazilian scenario, following the guidelines of descriptive research.

It has a predominantly qualitative approach, more appropriate to the social studies as stated by Richardson (2008). The depth of qualitative analysis is justified in so far that the study transcends usage verification. However, has collection of quantitative data, where quantitative data are analyzed qualitatively (Creswell 2007).

The research objects are the digital libraries linked to the government agency, in which the variables are the descriptive elements. Thus, the used elements and qualifiers of Dublin Core are accounted, so it is possible to compare and analyze the results.

3. Results

The study identified 13 digital libraries linked directly with Brazilian government agencies, all designed with DSpace, as shown in Table 1, providing more than 427,000 documents in full text. Thus, there are four libraries from the executive power, five from the judicial and four of the legislative branch. This reveals the interest of the Brazilian government agencies in the use of DSpace, which was developed primarily for the development of academic systems. The Digital Library of Housing (Biblioteca Digital da Habitação - HABI) from São Paulo is fully restricted, preventing outside access to their documents, so staying out of the research.

TABLE 1 – List of analysed digital libraries

Branch	Agency Government	Library Name	Records	URL
Executive	Ministério do Planejamento, Orçamento e Gestão	SPI - Biblioteca Digital do Planejamento	494	http://bibspi.planejamento.gov.br
Executive	Ministério Público Federal - MPF	Biblioteca Digital do MPF	21.335	http://bibliotecadigital.mpf.mp.br/xmlui
Executive	Secretaria Geral da Presidência da República	Biblioteca Digital da Participação Social	395	https://biblioteca.participa.br/jspui
Executive	Prefeitura de São Paulo	Biblioteca HABI	11.322	http://biblioteca.habisp.inf.br
Judiciary	Tribunal Regional Federal da 1ª Região	Biblioteca Digital TRF1	44.977	http://www.trf1.jus.br/dspace

Judiciary	Superior Tribunal de Justiça - STJ	Biblioteca Digital Jurídica -STJ	76.124	http://bdjur.stj.jus.br
Judiciary	Tribunal de Contas do Município do Rio de Janeiro	Biblioteca Virtual em Controle Externo	150	http://bvce.tcm.rj.gov.br
Judiciary	Tribunal de Justiça do Estado do Ceará - TJCE	Biblioteca Digital Jurídica -TJCE	502	http://bdjur.tjce.jus.br/jspui/
Judiciary	Tribunal Superior do Trabalho - TST	Biblioteca Digital do Tribunal Superior do Trabalho	8.626	http://aplicacao.tst.jus.br/dspace
Legislative	Senado Federal	Biblioteca Digital do Senado Federal	262.210	http://www2.senado.leg.br/bdsf
Legislative	Câmara dos Deputados	Biblioteca Digital da Câmara dos Deputados	3.516	http://bd.camara.leg.br/bd/
Legislative	Câmara Legislativa do Distrito Federal	Biblioteca Digital da Câmara Legislativa do Distrito Federal	48	http://biblioteca.cl.df.gov.br/dspace
Legislative	Assembléia de Minas	Biblioteca Digital da ALMG	13.372	http://dspace.almg.gov.br/xmlui

A point to note is that out of the 13 libraries selected for analysis, only three provide interoperability via Open Archives Initiative - Protocol Metadata Harvesting (OAI-PMH), even if using DSpace software, a system where this option is very easy to implement. This indicates poor adherence to the precepts of open files, an political issue, as these institutions do not have the same concern with institutional visibility as academic institutions have.

The poor adherence to interoperability by government repositories can be explained by the absence of a federation to join all these repositories, such federation could offer services such as consolidated searches on government digital documents. Thus, it requires that repositories make available the OAI-PMH in order to establish interoperability, revealing certain isolation between government repositories.

Regarding the executive branch, it proves to be present in the various levels of public action, with repositories linked to the Presidency library up to city halls. Emphasis is on the Digital Library of Social Participation, created in 2014, linked to the General Secretariat of the Republic Presidency (Secretaria Geral da Presidência da República), focused on the dissemination of government documents on social participation in government actions. This digital library is linked to the higher Brazilian administrative level.

The judiciary has the highest amount of digital libraries, at the various hierarchical levels of power. Emphasis on the Digital Library Legal (Biblioteca Digital Jurídica), developed by the Superior Court of Justice, being the first nonacademic Brazilian institution to make use of DSpace for creating an information system, in operation since 2005. This library has stimulated the use of DSpace in other legal institutions, with support from the Brazilian Institute of Information Science and Technology (IBICT).

The Legislative, in turn, has digital libraries in the Senate and the House of Representatives, revealing the adherence of this tool by higher levels of legislative organs. In the Library of the Senate highlights the collection of articles in newspapers and magazines, in order to preserve its institution memory through this documentation. In the Digital Library of the Federal Chamber, highlights are to the historical documents of the Brazilian Republic. Together these two libraries provide over 250 thousand documents.

In digital libraries developed with DSpace, the classification process is presented in the form of organizing the collection in communities, sub-communities and collections. At this point, it appears that most digital libraries categorize documents by document type (six libraries), followed by the organ activities (two libraries) and to the organizational structure of the agency (one library). There are libraries that present joint document type/organizational structure categorizations (three libraries). There is no standardized form of collection organization, with only one digital library, the Biblioteca Digital da Participação Social, organized by thematic taxonomy of the organ.

Even with minor variations, all repositories use the qualified Dublin Core metadata schema, despite DSpace's flexibility to use other schemes, which shows the adaptability of Dublin Core to describe a variety of document types. There was a wide variation in the use of metadata, not only on the amount used, but also on the elements and qualifiers. The Digital Library of the Regional Court of Ceará (Biblioteca Digital Jurídica - TJCE), for example, uses only 11 different metadata fields to describe the documents, while the Digital Library of the Superior Labor Court uses 43. This variation reveals little standardization in the description of the documents, as these two libraries are from the judiciary branch.

The most commonly used elements in all libraries are dc:contributor, dc:date, dc:identifier and dc:title (Figure 1). The Digital Library of the Superior Labor Court adds elements and qualifiers Electronic Thesis and Dissertation - Metadata Standard (ETD-MS), as it contains theses and dissertations in its collection. The Biblioteca Digital do Tribunal Superior do Trabalho created an element called dc:atos, to contain the identification of documents called "act", the only new element identified.

The research has revealed that dc:description is the metadata element used with more different qualifiers in these repositories. This can be explained because when there isn't a specific element to describe a digital object characteristic, many repositories' managers use the flexibility of this element on the description. Also, element dc:date is used the same way, as there are a lot of dates to describe a digital object, like creation date, submission date, publication date, and so on. Another point is about dc:element identifier, usually a digital object has a unique identifier but in repositories there can be noticed two identifiers, URL and digital object own identifier.

This findings contrast in part with Alijani and Jowkar's (2008) research results, highlighting the differences between academic digital objects and governmental digital objects. In fact for academic digital objects, title element is very important, but in governmental documents the description is as important as the title, as far as in some cases governmental digital objects's title is sometimes irrelevant.

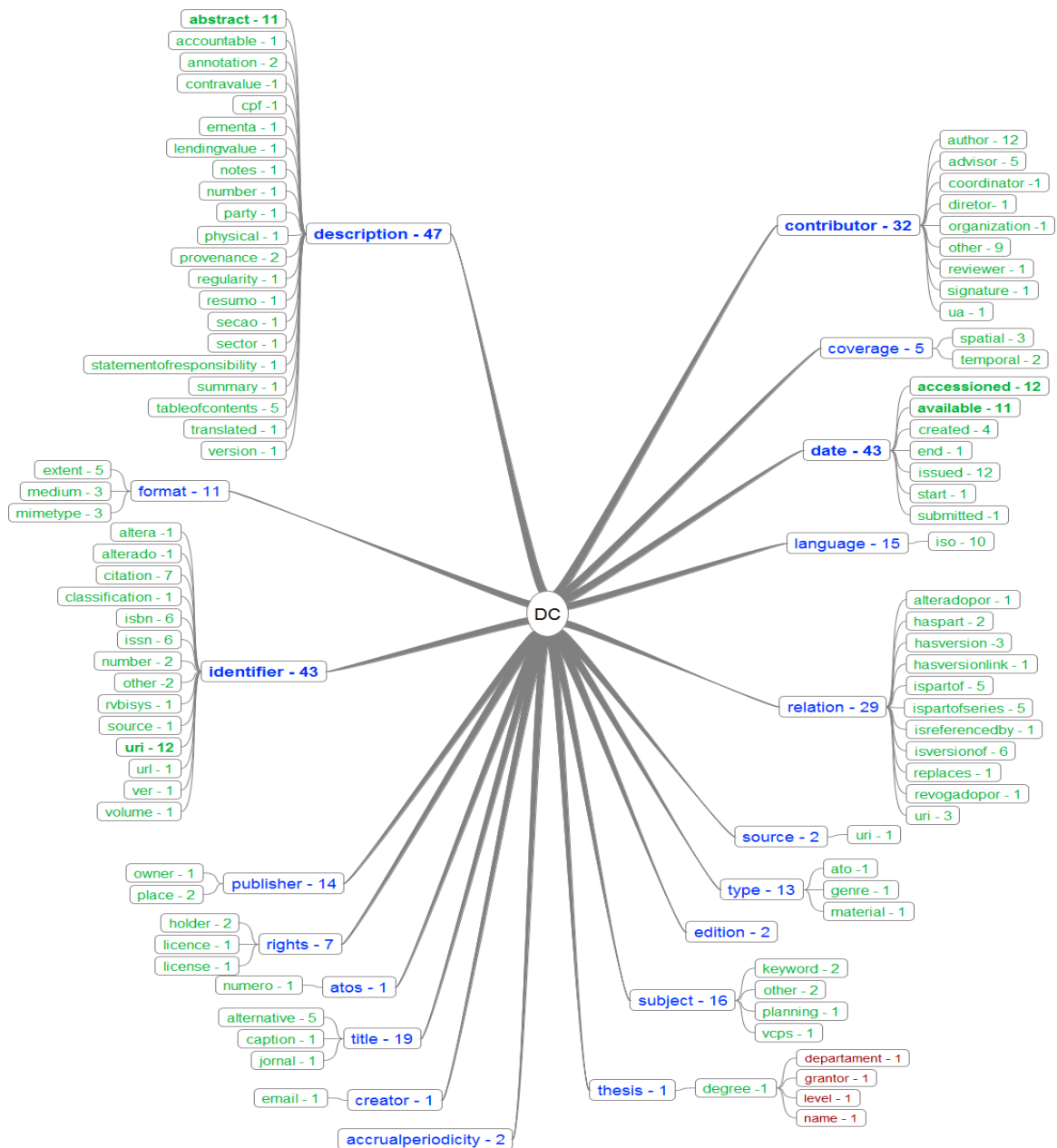


FIG. 1. Number of times that the elements and qualifiers are represented.

All digital libraries are using the author field (dc.contributor.author), title (dc.title) and date of publication (dc.date.issued), being these the most frequent, followed by summary (dc.description.abstract) and editor (dc.publisher), that does not appear in one library, the Biblioteca Digital TRF1. Another point is that 43 metadata fields are used by only one institution, the Tribunal Superior do Trabalho, indicating low standardization or specific needs to describe its documents.

The wide range of qualifiers can be highlighted in Figure 2, which presents the use of qualifiers per element. Noteworthy is the large number of qualifiers of elements dc:description, dc:identifier, dc:relation; dc:contributor and dc:date, noting that the description of government documents takes place in these elements.

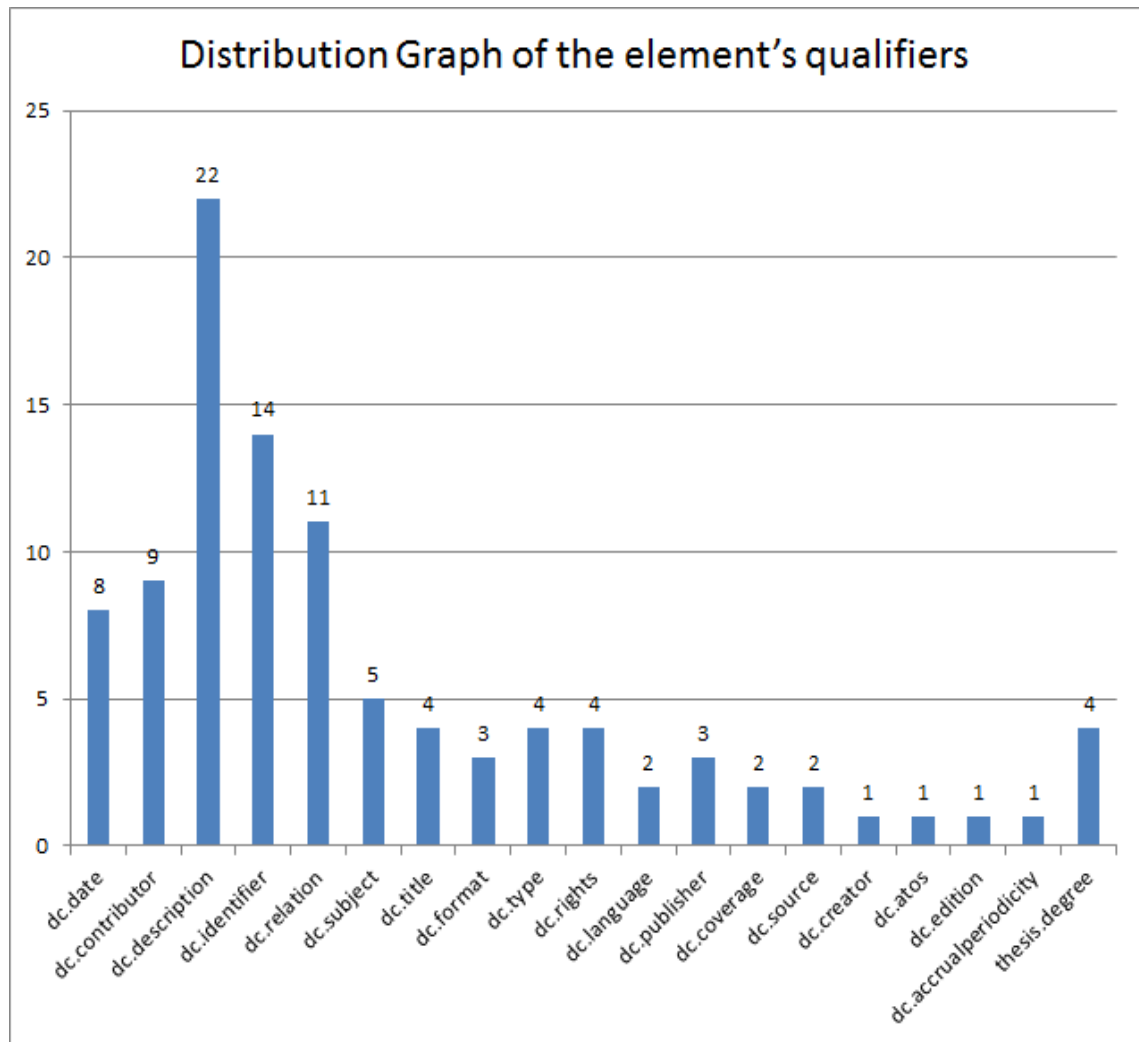


FIG. 2. Distribution Graph of Element's Qualifiers

As for the qualifiers of elements can be highlighted:

- The use of dc:description to describe the characteristics of documents that do not have elements provided in DC, in many cases, creating qualifiers to the description;
- As government documents have specific identifications, the large amount of dc:identifier qualifiers;
- The use dc:relation to indicate the various types of relationship between government documents;
- Government documents have lots of contributors, so lots of qualifiers;
- Dates are important in government documents, so the large number of qualifiers in dc:date.

This shows specific requirements of government documents in front of traditional academic digital libraries, even though in a few cases certain discrepancy in the understanding of elements, qualifiers and its content. However, as interoperability is not a concern on these libraries, this is not a big problem.

3. Final Remarks

The results revealed interest of government institutions on DSpace in the construction of government digital libraries, in part by the action of IBICT for the dissemination and support offered to user's community, even for non-academic institutions.

Also, the study found that government libraries do not use taxonomies related to its area to organize documents, preferring to use document types or organizational structure. As not offering interoperability there is not a concern with standardization of metadata fields, making use of wide variation due to the documentary specificity.

Also arise perspectives for the study of government documents classification in digital libraries and repositories, in order to facilitate its organization and retrieval, using government related taxonomies for example, and supporting the organization of this type of documents on the web.

The use of the dc:description elements can be observed to adapt Dublin Core to describe the government documents, revealing the flexibility of this metadata schema for describing a varied documentary typology. This point may be evidence of the need for studies for the proposal of more specific elements or qualifiers for these type of documents in the context of Brazilian government.

In addition, an analysis of the users of government repositories, their expectations, experiences and requirements regarding what they seek in the repositories can guide the planning and preparation of metadata application profile.

An increased number of libraries, archives or other initiatives on the Internet using the Dublin Core to describe the documents present challenges and opportunities studies. In Brazil, this is a promising scenario as a recommendation, all government documents must be accessible ads defined by the Information Access Act, No. 12,527, from November 18, 2011.

References

- Alijania, A. S.; & JOWKAR, A. (2009). Dublin Core metadata element set usage in national libraries' web sites. *The Electronic Library*, 27(3), 441-447.
- Creswell, J. W (2007). *Projeto de pesquisa: método qualitativo, quantitativo e misto*. Tradução de Luciana de Oliveira da Rocha. 2. ed. Porto Alegre: Artmed.
- Gil, A. C (2006). *Como elaborar um projeto de pesquisa*. 4. ed. São Paulo: Atlas.
- Richardson, J. R (2008). *Pesquisa social: métodos e técnicas*. 3. ed. São Paulo: Atlas.

BEAM Repository: A Proposal for Family and Personal Repository

Rachel Cristina Vesu Alves
São Paulo State University,
Brazil
rachel@marilia.unesp.br

Ana Carolina Simionato
Federal University of São
Carlos, Brazil
acsimionato@ufscar.br

Felipe Augusto Arakaki
São Paulo State
University, Brazil
fe.arakaki@marilia.unesp.br

Paula Regina Ventura Amorim
Gonçalves
São Paulo State University,
Brazil
paulagoncalvez@marilia.unesp.br

Ana Paula Grisoto
São Paulo State
University, Brazil
grisotoana@reitoria.unesp.br

Plácida Leopoldina
Ventura Amorim da Costa
Santos
São Paulo State
University, Brazil
placida@marilia.unesp.br

Abstract

Preservation of cultural heritage has been widely discussed in the last decades. Different groups of people contribute to the production and preservation of cultural heritage through personal and family performance. However, there is a lack of environments specifically prepared to store and organize the resources produced by these groups, resulting in difficulties to access and preserve these materials along the time. The hypothesis is that the digital repository and the structured metadata standards are relevant tools to provide the suitable environment to store, describe, access and preserve family and personal resources. The study herein has a theoretical and applied basis, for it aims to investigate and confirm the hypothesis using theories and applying them. It aims at demonstrating that the digital repositories are relevant for the storage, description, access and preservation of personal and family information. During implementation of the digital repository, DSpace software and Dublin Core standard were used. As a result, the implemented repository showed itself as a viable alternative for storing this information. It is possible to conclude that such a digital repository constitutes a tool that guarantees the preservation, access and sharing of archives, resources and data produced by families and individuals in the digital environment.

Keywords: BEAM Repository; DSpace; Dublin Core; Family and personal repository.

1. Introduction

Preservation of cultural heritage has been widely discussed in the last decades. Many institutions are providing their preserved cultural patrimony through their digital collection. In this scenario, the big challenge is to provide the suitable representation of digital information resources, guaranteeing the integration of different communities and the interoperability of data. Thus, the use of metadata and metadata standards has become a common practice among the several areas that seek to preserve and provide cultural heritage in digital collection.

Cultural patrimony consists of several categories like the tangible cultural patrimony (paintings, sculptures, manuscripts, monuments, cities, shipwrecks, ruins, etc.), immaterial cultural patrimony (oral traditions, arts, music, etc.) and the natural patrimony (natural reservation, archeological or geological sites, etc.) (UNESCO, 2009). Thus, different groups of people, especially individuals and families, contribute to the production and preservation of cultural heritage, building and perpetuating the tangible cultural patrimony, immaterial cultural patrimony and the natural patrimony.

Ordinary people and their family can find on the Internet environments such as Facebook, Flickr, Blogs, Instagram, which provide access to some of their personal information. However, people provide a diversity of family and personal content such as pictures, documents, videos, intellectual and artistic productions, material related to trips among others, and they do not have a specific environment to store and organize these resources options, resulting in difficulties to access and preserve these materials along the time.

The personal and family contents, denominated herein as personal and family information resources, in many cases held by one or more members of the family, are sometimes discarded or do not receive informational treatment that guarantees the access and makes them easy to be located. Thus, it is important to provide storage, adequate description of resources, preservation and, at the same time, extend the access and the sharing of personal or family production, reducing physical and temporal spaces among people from the same family core. Such actions contribute to preserve cultural heritage produced inside a family environment or by the personal interaction with the existing cultural heritage.

In the last years, a growing number of digital repositories were developed by different types of organizations such as digital libraries, universities, public archives, and research centers among others. The Open Archives Initiative and the creation of open source softwares's made it easier to provide digital contents of these organizations. However, it is possible to see that there are not many initiatives of digital repository implementation with familiar and personal purpose.

Digital repositories are considered environments that provide storage, description, organization and preservation of information resources, guaranteeing that their access and the family or personal history is not lost along the time.

The hypothesis for this study is that the digital repositories and the structured metadata standard are relevant tools to provide an environment able to store, preserve and provide the access to family and personal information resources in a more organized way.

This study has an applied basis, for it aims to investigate and confirm the hypothesis established and solve practical and immediate application problems. It is also considered a qualitative and exploratory study because it seeks information in order to clarify the subject investigated taking into account several aspects (Cervo & Bervian, 2003; Gil, 2002).

The aim of this study is to demonstrate, by means of an implementation, that the digital repositories constitute themselves as relevant environment to store, describe, access and preserve the personal and family information resources. It is possible to conclude that the implementation of such a digital repository constitutes a tool to guarantee the preservation, access and sharing of collection of resources archives and data produced by family cores and individuals in the digital environment.

2. Digital repository and metadata

According to Pollak (1992, p. 204), memory contributes to build individual and collective identity. The family and personal information resources represent the bond between individuals and their lives, performing an important role in registering and perpetuating memory, for they bring memories of previous experiences, places they visited, their ancestry and their life history.

With the advance of technologies, several kinds of family and personal information resources have started to be produced, requesting organization to be available in digital environment.

In doing so, the implementation of digital repositories is presented as a relevant initiative to store, share and access resources. Repositories can be described as

systems available in the web that provide, mainly, facilities to add and access digital objects . . . repositories aggregate a great variety of facilities, most of them related to management of digital objects added . . . besides managing digital documents, they have facilities related to their preservation and they are flexible systems that can adequate themselves to fit several purposes (Shintaku & Meireles, 2010, p. 17).

There are different repositories such as academic, administrative, technical and hybrid. According to a more general classification, there are also institutional and thematic repositories. Regarding the thematic repositories, the object of this study, it is possible to highlight that the origin of the information resources provided is diverse and that the theme is the main point of aggregation of these resources, making their access easier (Shintaku & Meireles, 2010). In personal repositories from Brazilian initiatives, information resources are organized by grouping the kinds of resources.

In this study, the purpose is to keep and manage resources for a long period and provide the appropriate sharing and the access to those interested in a thematic structure. The software chosen for implementing the repository was DSpace¹, because according to the Registry of Open Access Repository (ROAR)² it is the most used software for implementing open source digital repositories. This software meets most requirements listed in open-source³ software analysis in the literature and it is widely used.

DSpace is open-source software to store, manage and distribute collections in digital format. The software meets the necessary requirements for the implementation of the repository, which is the purpose of this study: open source software. It does not have costs to be acquired and has simple and intuitive web interface; stores different kinds of information resources. It allows the inclusion of more than one format of archive per work described and the creation of distinct collections. It manages collections of items, communities and sub communities with more than one collection; establishes relations among resources, collections and their preservation; stores, imports and exports resources and their metadata, according to Dublin Core standard or other standard if necessary.

It supports many idioms in the metadata and digital content field; uses unique identifiers (Handle System); provides digital preservation compatible with Open Archive Information System model (OAIS); shares metadata through protocol Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), exports archives in Metadata Encoding and Transmission Standard eXtensible Markup Language (METS XML) and also works with Open Uniform Resource Locator (URL) protocol; it presents levels of customization in both user interface and its structure; offers access control to communities, sub communities, collections and resources by means of determining activities. Interaction with user is performed by email and with available information in the repository; the interface with the user allows browsing among communities, sub communities and collections. Browsing and searching can be performed by creator, title, subject, date and key-words found in metadata; the submission can be performed by the creator of resources; it provides a satisfactory documentation for its implementation (Pirounakis & Nikolaidou, 2009; Sayão & Marcondes, 2009; Romani, Fusco & Santos, 2010).

DSpace suits better in cases when it is necessary to establish communities, sub communities and collections, manage information resources and submit these resources. These are determinant characteristics for the repository at issue, which is in process of implementation inside the Library of Study and Application of Metadata (Biblioteca de Estudos e Aplicação de Metadados - BEAM). Another fundamental characteristic was the possibility of using Dublin Core standard for the description of information resources, for this is a metadata standard, which can be used by experts or those who are not experts. The aim is that any individual, expert or not, can organize their family and personal information resources in a digital environment. The idea is using a system in which the following functionalities are observed: a) easy collection and insertion of information resources, including their metadata; b) easy access to information resources either by list of communities, sub communities, collections and items or also by the search interface; c)

¹ Retrieved March 25, 2015, from <http://dspace.org/>.

² Retrieved March 25, 2015, from <http://roar.eprints.org/view/software/>.

³ "software package whose distribution follows its source code, allowing the user to modify and adequate the software according to his necessities" (Toutain, 2006, p. 20).

promotion of long-term preservation of information resources stored in this system (Lewis & Yates, 2008).

The metadata standard adopted to implement and implant the repository herein is the Dublin Core developed from characteristics such as simplicity; semantic interoperability; international agreement; extensibility and modularity of metadata in the Web. Such characteristics are understood as follow: the proposal of simplicity makes it possible for those who are not experts to describe a resource in the Web; the semantic interoperability comprehending several metadata standards enabling the interoperability; the agreement in accepting internationality the Dublin Core Metadata Element Set; the extensibility and flexibility of Dublin Core when widening and adding descriptive elements, which are presented as elective and repetitive; the modularity of metadata in the Web, metadata can be combined with other schemes, even if they contemplate different semantic and syntactic structures (Alves & Santos, 2013).

3. Development and Implementation of BEAM Repository

BEAM Repository⁴ is a project created by the BEAM, of the Group of Research in New Technologies in Information (Grupo de Pesquisa Novas Tecnologias em Informação – GPNTI), in Philosophy and Science University at São Paulo State University (UNESP), Marília/SP. BEAM aims at providing the students linked to the library of studies and application an environment for the development of researches related to the creation and manipulation of digital objects metadata. In doing so, the goals in constructing the repository are: a) to provide an environment for studies and practical applications over metadata and metadata standards; b) to manage digital collections; c) to make possible the study and practices relating to the interoperability, harvesting, digital ownership among others.

The first initiative using the implementation of BEAM repository was the creation of family and personal archives to organize travel material, in order to make the access to information resources easier, preserving the history and memory of the individuals in the family.

The implementation of BEAM Repository was performed based on the following phases:

- Phase 1 – Planning and defining the repository scope: it started after identifying the necessity of organizing, in a digital way, travelling materials produced or acquired by families and people. Later, actions to solve the problem were taken, contributing to the establishment of the repository scope to be implemented and to the adoption of Dublin Core metadata standard. These actions were distributed in a more detailed way in the following phases. The software was also chosen according to what was mentioned in item 2 herein.
- Phase 2 – Implementation and personalization of software: this phase embraced the installation of DSpace software in the research group server, the personalization and configuration of its visual interface (model, layout, sources, colors, BEAM logo insertion etc.).
- Phase 3 – Definition of metadata in the repository: the simple scheme of Dublin Core metadata standard was chosen in the planning phase so that the template used to describe the information resources at the moment of their insertion in the repository could be built. This phase is related to the previous phase, because it is also related to the personalization of DSpace software.
- Phase 4 – Definition of communities and sub communities: in order to insert information resources in the repository, it was necessary to establish first the communities and sub communities that would group the collection of information resources. The collection was classified according to the primary and secondary needs. The primary needs were defined based on the family composition, that is, the family is composed of two or more people related by birth, marriage, adoption, civil union or some other similar legal way that groups a

⁴ <http://beam.marilia.unesp.br>

family (International Federation of Library Associations, 2009). The secondary necessities are related to people belonging to a family. Therefore, person is defined as an individual or identity established individually or in group (International Federation of Library Associations, 2009). This way, the superior hierarchic position corresponds to the necessities of family grouping and the subordinated positions correspond to people belonging to the family. Thus, the communities will be families and sub communities, and the people related to these families, with the possibility of having different hierarchic structures.

- Phase 5 – Definition of collections: the definition of the collections in communities or sub communities was established from the idea of events that represent an action or occurrence with these people and families (International Federation of Library Associations, 2009). This way, families and people can organize their digital resources in a diversity of collections that correspond to several kinds of events or occurrences (travels, birthdays, weddings, vacation etc.). The collections can be established individually or they can be related to other people or even to the family.
- Phase 6 - Definition of search system: the DSpace software provides three kinds of search to discover and recover resources: surfing the communities, sub communities and collections; the simple and advanced search, which can be refined by title, author, subject, date of publication (of the information resource in the repository); and date.
- Phase 7 - Definition of a Use guide: the definition of the user guide was established considering the lay user who can access the metadata scheme for inserting information resources in his collection. It is a case of defining metadata, guidelines for entering the value and defining which metadata is considered obligatory for the scope of this repository. Although all Dublin Core standard metadata is optional, it was necessary to establish some obligatory metadata so that the repository and its search system could work minimally. The information in the user guide was inserted in the repository description template so that the users could access the information at the moment of the description, without the necessity of accessing other resources.
- Phase 8 – Insertion and description of resources in the system: the information resources of the fictitious family were inserted in communities, sub communities and collections, based on the template description. Some resources were already in digital form and the printed resources were previously converted to digital form to be inserted.

The BEAM repository has already gone through all the phases of its development and implementation and currently it is in a phase of evaluation and performance of some adjustments considered necessary to a good operation of the repository. One of these adjustments is related to the advanced search system settings, which are being improved in order to include the type filter (kind of resource) and correct the date filter (dc. date). It would make it even easier for people from the same family to locate and access the resources.

After performing all the phases, some considerations about the practical application of the repository are presented as follow. First, in relation to the BEAM repository, it is possible to observe that it corresponding to the DSpace hierarchic structure through which the information resources are released: Communities, Sub communities, Collections and finally the Items. The Communities are composed from the union of two people, forming a family that conducts the relationship of the other members.

This family can have joint events called Community Collection, which have joint information resources. Sub communities are members of this family and incorporate resources and individual occurrence called Events (Collection).

In doing so, figure 1 hierarchically shows the family, the person and the event, composing respectively the communities, sub communities and collections structured in BEAM Repository.

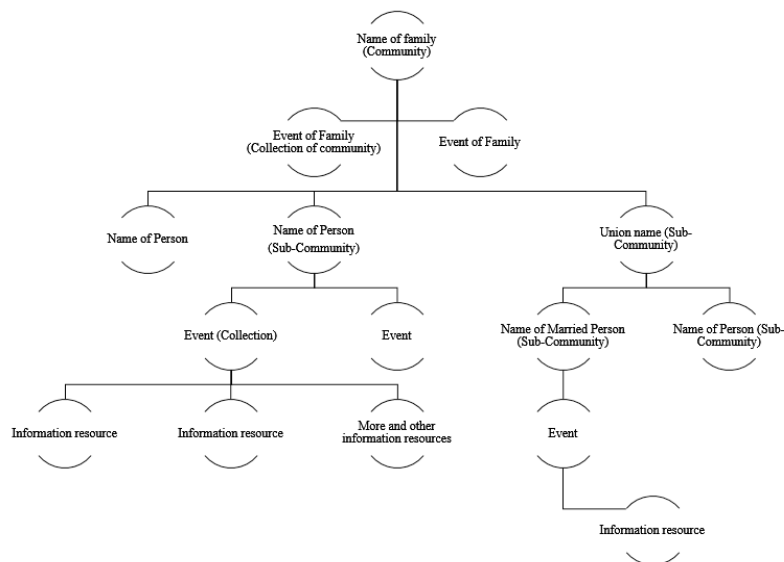


Fig. 1. Hierarchy built for family and personal archives

From this hierarchy, a fictitious family community was created, the Silva's family. The patriarch José Silva is a Portuguese immigrant, from Porto city and arrived in Brazil in 1950. He met Maria Souza in Santos, São Paulo, they got married in 1959 and had three children: José Silva Filho, Paula Silva and Carolina Silva. One of the daughters, Paula, got married to Luiz Oliveira, creating a new family related to the first one. The names of the family members are communities and sub communities and the collections can be private or with other members of the family.

Figure 2 shows the hierarchy built. The fictitious names appear in alphabetic order on the repository page.



BEAM

Repositório BEAM

[Página inicial](#) → [Lista da comunidade](#)

Buscar DSpace

Ir

[Busca avançada](#)

Navegar

Todo o repositório

[Comunidades e Coleções](#)

[Por data do documento](#)

[Autores](#)

[Títulos](#)

[Assuntos](#)

Minha conta

[Sair](#)

[Perfil](#)

[Submissões](#)

Contexto

[Criar comunidade](#)

Administrativo

[Painel de controle](#)

Controle de acesso

[People](#)

[Grupos](#)

Comunidades no DSpace

Selecione uma comunidade para navegar nas coleções.

- [Família Silva: José e Maria](#)
 - [Viagem Porto, Portugal - 2013](#)
 - [Carolina Silva](#)
 - [Viagem Florença, Itália - 2008](#)
 - [Viagem Maranello, Itália - 2008](#)
 - [José Silva](#)
 - [Viagem Alemanha - 2008](#)
 - [Viagem Salvador, Bahia - 2011](#)
 - [José Silva Filho](#)
 - [Viagem Espanha - 2006](#)
 - [Viagem São Paulo - Março 2015](#)
 - [Maria Souza Silva](#)
 - [Viagem Rio de Janeiro - Outubro 2012](#)
 - [Viagem Viena, Áustria - 2011](#)
 - [Paula Silva Oliveira e Luiz Oliveira](#)
 - [Felipe Silva Oliveira](#)
 - [Viagem Viena, Áustria - 2014](#)
 - [Luiz Oliveira](#)
 - [Viagem para Paranaguá, PR](#)
 - [Paula Silva Oliveira](#)
 - [Viagem para Ilha do Mel - PR](#)

Fig. 2. Silva's family

The information resources which can be inserted in the family and personal repository, were defined as follow: images, maps, slides presentations, music recordings, sound recordings, travel brochures and leaflets, touristic guides, museum and subway tickets, air tickets, videos, travel cards, stamps and postcards, personal documents such as certificates, letters among others. Other types of information resources can be inserted. The repository holder defines the criteria from the considerations that take into account the level of importance of resources for preservation, memory, guard and access with the community to be attended.

Resources description is performed with simple Dublin Core, with its 15 description elements, because it is a suitable standard for users who are not experts in generate metadata.

Considering that the user of the system is not an expert, there was the necessity of adapting the representation of information (metadata values). To do so, metadata guidelines and definitions were proposed and obligatory and optional metadata was indicated for this repository, as follow:

- Title: (obligatory element) main title of the resource or title in which it can be known. Insert preferably the title in the resource. In case you do not find it, attribute a title;
- Creator: (optional element) insert the resource creator. Assign the responsibility for creating the resource to the person or group that has more intellectual or artistic responsibility. Some possibilities are the author, editor, photographer, producer among others;
- Subject: (obligatory element) introduce the subject or keyword, which represents the resource. It is recommended to insert at least three subjects in order to make the recovery easier;
- Description: (optional element) insert any suitable description or comments to represent characteristics in relation to the resource: characteristics of the place visited, characteristics about the document stored or personal comments. It is recommended to insert information about time and day of visit or the contact information about the place visited; permission for registering images or videos in the place; permission to use flash; information about people related to the information resource or any other information considered relevant to the resource described herein;
- Publisher: (optional element) insert the name of who published the resource described. The publisher is the person or group of people responsible for publishing and distributing the material;
- Contributor: (optional element) insert the name of those who contribute with the resource. It can be a person or group of people that contributed intellectually or artistically for the creation of information resource. However, this person is not the main responsible for the resource;
- Date: (optional element) insert a date or period of time referring to the resource. It can be a date when the resource was acquired, the date when the image was obtained, date when the photo was taken, date when a video was recorded, etc. It is recommended that the dates are standardized according to what was established by W3C based on ISO 8601 (for example YYYY-MM-DD). In case there is no accuracy, use approximate month and year or just the year.
- Type: (optional element) insert the type (kind or nature) of the resource. Select the type of resource according to the values in the list (to select more than one value, keep the "Ctrl" or "Shift" keys pressed);
- Format: (optional element) insert if the format is physical or digital. Insert the file format, physical environment or resource dimensions;
- Identifier: (optional element) insert an identifier or a unique reference for the resource (item) in a context;

- Source: (optional element) insert the source or origin in which the resource was derivated. It is a case of indicating the original source that derivated the resource, or the relationship between the parts of a resource. It is recommended to insert the name of the person who has the original resource ;
- Language: (optional element) select in the list the main idiom of the resource. In case it does not appear in the list, select "Other" and if the resource does not present an idiom, like in the case of the images and photographs, select "N/A";
- Relation: (optional element) insert the resource relationships. The relationship indicates if a resource is a physical or a logical part of another resource, if it is a version of another resource, if it presents a transformation, if it is a reproduction, etc.,
- Coverage: (optional element) insert a coverage: spatial location, temporal period or a jurisdiction referring to the resource. For the spatial location, insert the name of the city and country; for the time coverage, insert a period of time or dates interval; for the jurisdiction, insert the name of the jurisdiction;
- Rights: (optional element) insert information concerning the resource rights. It includes the declaration of rights about the access and availability of resource, the indication of intellectual property rights, copyrights, etc.

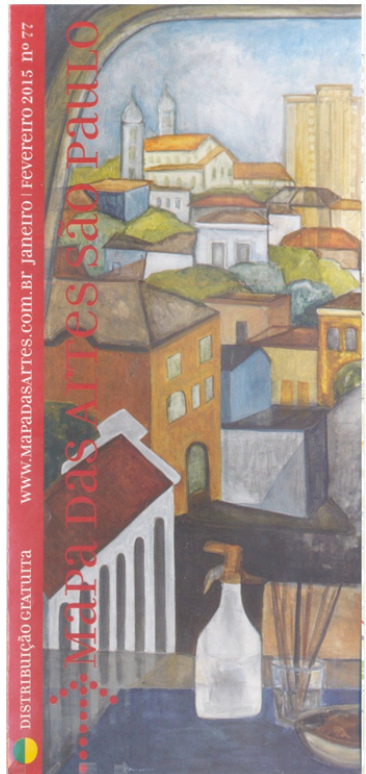
BEAM repository has defined the control of access to the communities, sub communities, collections and resources, by means of determining permissions. Communities, sub communities, distinct collections and management of collections and items already implanted.

Silva's Family community, which is the example herein, is already available to be used in 45 resources Web, in different ways: travel brochures, photographs, videos, maps, tickets, visit guide, among others. In BEAM repository, the definition of relations between resources and collections is in process of validation.

The proposal is that in the case of the family repository, the interaction is performed through information available in the repository, with an interface that allows the browsing between communities, sub communities and collections.

The main concern in developing the repository is to guarantee facility in collecting, inserting information resources and determining metadata values; and in accessing information resources, either through list of communities, sub communities and items or through search interface. The idea is offering, in digital format, the leisure possibilities provided by family albums in gathering resources that tell a little about the characteristics of a family history or of a person in a pleasant and easy access.

Dublin Core standard contributes to this purpose because it is easy to understand and it is constituted of few metadata, generating a disposition to feed descriptive values. In doing so, the description of "São Paulo arts map" resource (Figure 3), the type of map, belonging to José Silva Filho sub community, is presented as a way to illustrate what was developed.



dc.contributor	Design Source	en
dc.creator	Celso Fioravante	en
dc.date	2015-03	
dc.date.accessioned	2015-03-28T14:55:58Z	
dc.date.available	2015-03-28T14:55:58Z	
dc.identifier.uri	http://beam.maria.unesp.br/xmlui/handle/123456789/79	
dc.description	The map offers exhibitions and artistic cultural points of São Paulo. The period covered by the exhibitions is from January to February 2015. The cover of the map is beautiful for that reason was saved.	en
dc.format	jpeg	
dc.format	pdf	en
dc.language.iso	pt_BR	
dc.publisher	Pancrom Graphics	en
dc.relation	http://www.mapadasartes.com.br/	en
dc.rights	openAccess	en
dc.source	José Silva Filho	en
dc.subject	art galleries	en
dc.subject	galleries	en
dc.subject	art	en
dc.title	São Paulo Map of Arts	en
dc.title	Mapa das Artes de São Paulo	
dc.type	Map	en
dc.coverage	São Paulo (capital), Brazil	en

FIG. 3. Register of São Paulo Arts Map

It is interesting to observe that a proposal of family and personal repository requires that the description of events and resources go beyond the descriptions defined formally by schemes, codes and standards of the area. It is necessary, in such work, an informal action in describing the resource stored, a description that refers to the proposal of preserving the family history in which there are notes of recommendation, expression of feelings about moments experienced in life, description about a special family moment. The way to show this was a concern in constructing the repository.

The alternative was using the dc.description option, because it satisfactorily receives values that are not controlled in the description of a resource whose purpose is, somehow, to preserve an affective and emotional bond in the preservation of personal history.

5. Final Considerations

Information resources are gathered during travels and individual or collective experiences and many times they do not have an easy access and organization. In order that these resources are not lost in space and time, it is important to create an environment that allows the access, recovery, sharing, use and reuse of these resources and that also allows the preservation of a community memory.

The aim of this study was to present an environment that provides an organization of family and personal digital information resource.

The repository implementation using DSpace had good results, although its installation and personalization demands specific knowledge. After implemented, DSpace software efficiently meets the structuring and representation requirements of the family and personal digital information resource.

In relation to the storage and description of information resources, it was possible to see that simple Dublin Core standard sufficiently meets the expectations and necessities of representing the information resources inserted. As it is a metadata standard dedicated to experts and to those

who are not experts, the description of resources can be made by a professional or by the members of the community attended.

The proposal is considered innovative not only from the point of view of the librarian, but also the kind of public attended. A family and personal repository is believed to be a viable alternative for storing information resources, guaranteeing memory preservation and family history construction, besides organizing the representation and preservation of information resources and family and personal data.

In BEAM, the development of manuals and tutorials for those who are not experts is in development as a continuation of this study, besides the continuation of studies referring to personalization of personal and family repository.

Acknowledgements

The authors would like to thank Jorge Janaite Neto and Luiz Felipe Galeffi contributions during the implementation and implantation of BEAM Repository.

References

- Alves, Rachel Cristina Vesu, and Plácida Leopoldina Ventura Amorim da Costa Santos. (2013). Metadados no domínio bibliográfico [Metadata in the Library Domain]. Rio de Janeiro: Intertexto.
- Cervo, Amado L., and Pedro A. Bervian. (2003). Metodologia científica [Scientific Methodology]. São Paulo: Prentice Hall.
- Gil, Antônio Carlos. (2002). Como elaborar projetos de pesquisa [How to develop research projects]. São Paulo: Atlas.
- International Federation of Library Associations. (2009). Functional Requirements for Authority Data: a conceptual model. Washington: IFLA. Retrieved January 30, 2015, from http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf.
- Lewis, Stuart, and Chris Yates. (2008). The DSpace Course - Introduction to DSpace. Retrieved March 22, 2015, from <http://cadair.aber.ac.uk/dspace/bitstream/handle/2160/617/Module%20-%20An%20introduction%20to%20DSpace.pdf?sequence=8&isAllowed=y>.
- Pirounakis, George, and Mara Nikolaidou. (2009). Comparing Open Source Digital Library Software. Retrieved March 22, 2015, from <http://galaxy.hua.gr/~mara/publications/ideaDL09a.pdf>.
- Pollak, Michael (1992). Memória e identidade social [Memory and social identity]. Estudos Históricos, 5(10), 200-212.
- Romani, Lucas Salviano, Elvis Fusco, and Plácida Leopoldina Ventura Amorim da Costa. Santos. (2010). Análise e implantação de Repositório Digital utilizando Software Livre DSPACE [Digital Repository analysis and deployment using Free Software DSPACE]. Proceedings of the Simpósio Brasileiro de Sistemas de Informação – SBSI, 2010, 01-16. Retrieved March 25, 2015, from <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2010/0019.pdf>.
- Sayão, Luis Fernando, and Carlos Henrique Marcondes. (2009). Software livres para repositórios institucionais: alguns subsídios para a seleção [Free software for institutional repositories: some subsidies for selection]. In: Sayão, Luis Fernando, Lúcia Brandão Toutain, Flavia Garcia Rosa, Carlos Henrique Marcondes. (Eds.). Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação (pp. 9-22). Salvador, Brasil: EDUFBA. Retrieved March 25, 2015, from https://repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf.
- Shintaku, Milton, and Rodrigo Meirelles. (2010). Manual do DSPACE: administração de repositórios [DSPACE guide: repositories of administration]. Salvador: EDUFBA, 2010. Retrieved March 30, 2015, from <http://www.repositorio.ufba.br/ri/handle/ri/769>.
- Toutain, Lúcia Maria Batista Brandão. (2006). Biblioteca digital: definição de termos [Digital Library: Definition of Terms]. In: Marcondes, Carlos Henrique, Hélio Kuramoto, Lúcia Brandão Toutain, and Carlos Marcondes. (Eds.). Bibliotecas digitais: saberes e práticas (pp. 15-24). Salvador, Brasil: EDUFBA.
- UNESCO. (2009) What is meant by "cultural heritage"? Retrieved April 1, 2015, from <http://www.unesco.org/new/en/culture/themes/illicit-trafficking-of-cultural-property/unesco-database-of-national-cultural-heritage-laws/frequently-asked-questions/definition-of-the-cultural-heritage/>.

The Use of Application Profiles and Metadata Schemas by Digital Repositories: Findings from a Survey

Morgana Andrade
Programa Doutoral em Tecnologias e
Sistemas de Informação –
Universidade do Minho - Portugal
morgana@dsi.uminho.pt

Ana Alice Baptista
Algoritmi Research
Center – Universidade
do Minho - Portugal
anaalice@dsi.uminho.pt

Abstract

Shows the results of a survey by questionnaire sent to the managers of 2, 165 digital repositories registered at OpenDOAR. Its purpose was to identify the existence and the use of application profiles and related metadata schemas. Of this total, 431 questionnaires were filled. The survey enabled the identification of metadata application profiles, as well as schemas and metadata elements/properties used within these repositories. According to the results, the number of repositories that use or provide metadata application profiles is 13, which we consider as very low. The Dublin Core remains as the most commonly used metadata schema, followed by MARC 21, METS and MODS. The dataset that resulted from the survey is openly available at RepositóriUM, the institutional repository of the University of Minho

Keywords: application profile; metadata schema; scientific digital repositories

1. Introduction

Metadata or data about data (National Information Standards, 2004, pp. 1) may be associated with a wide range of information and be adopted for different purposes. Based on its content and purposes, a DR may have metadata elements/properties drawn from a single or from several metadata schemas simultaneously, which leads us to the concept of Metadata Application Profile (MAP) (Heery & Anderson, 2005, Heery & Patel, 2000).

The concept of MAP has been evolving through the years. It started as a specification of a “mix and match” of metadata elements drawn from several metadata schemas (Heery & Patel, 2000), to a more complex construct as defined by the Singapore Framework for Application Profiles (Nilsson, Baker, & Johnston, 2008). For this study, we used the concept as described by (National Information Standards Organization- NISO (2007) which states that a MAP specifies how elements from one or more metadata schemas combine and fit to describe a particular set of resources, stipulating what and how the elements are adopted for description. By favoring the understanding of an application metadata model and relating it to existing schemas and encoding schemes, MAPs favor interoperability especially if they are encoded in a widely used linked data language such as the Resource Description Framework (RDF).

According to Curado Malta & Baptista (2014), various communities are defining and using MAPs. As an example there is the Scholarly Work Application Profile (SWAP), developed in 2008 to provide a method for describing scholarly works, research papers or scholarly research texts in Eprints UK (DCMI Usage Board, 2009). Another example is the RIO XX, also targeted to the UK institutional repositories ("RIOXX...", 2014). Other MAPs have been developed for specific domains or for specific institutions. An example is The Virtual Open Access Agriculture & Aquaculture Repository (VOA3R) MAP from the Food and Agriculture Organization (FAO) (Diamantopoulos et al. 2011). In the context of digital libraries there is the DC-Library Application Profile, developed by the Dublin Core Metadata Initiative (DCMI) (Guenther, 2000).

In what regards metadata, DRs have at least one thing in common: the OAI-PMH protocol. This protocol uses the simple Dublin Core (DC) metadata schema, which implementation is known in the community as OAI-DC. Although simple DC is a very good cross-domain metadata schema, there is an increasing need for domain-specific metadata elements in order to provide means for better relationships among resources and more accurate searches and results at a global level (Bruce & Hillmann, 2004, Chan, 2005, Clayphan & Oldroyd, 2005, Heery & Anderson, 2005, Hillmann & Phipps, 2007). It is reasonable to expect that some of the existing DRs already use more metadata elements than the ones provided by OAI-DC, or even have MAPs clearly defined, but there are not up-to-date studies about this reality (Park & Tosaka, 2010).

The main goal of this study is to identify the current panorama of DRs in what regards the use of metadata elements, their schemas and the definition of MAPs. Therefore, this study is proposed to: a) check if the repositories have clearly defined application profiles and which; b) identify the adopted metadata schemas and elements; and c) relate adopted metadata schemas and elements with the type of DR.

2. Methodology

This research adopted the survey by questionnaire for which we used Survey Monkey. The sample was restricted to the DRs registered at The Directory of Open Access Repositories (OpenDOAR - <http://www.opendoar.org/>) until September 4, 2014. The data collection was performed from September 2014 until November 2014. We selected only repositories with registered email addresses, regardless of type and geographical location, which corresponded to 2,165 repositories, out of a total of 2,720. OpenDOAR was selected because it has been widely used by the DRs community and European projects and initiatives, such as the Digital Repository Infrastructure Vision for European Research (DRIVER), the Surf Foundation and the Sherpa Services.

The questionnaire was structured in three sections, with a total of 11 questions. The first section aimed at the DR identification of the repository; the second section aimed at the verification of the existence of a MAP; and the third section aimed at the identification of schemas and metadata elements used by the DRs. For the sake of clarification, and to avoid misunderstandings, all the metadata related terms used in questions were properly defined before they were used.

In section 1, after the repository's name and/or acronym (question number 1 - Q1), we requested its type (question number 2 - Q2). Based on literature, we consider that an Institutional Repository (IR) stores the intellectual production of a research institution; a Thematic Repository (TR) stores domain-specific research results; an Organizational Repository (OR) stores documents/artifacts of an organization whose main aim is not related to research (e.g., the DR of the Brazilian Federal Court); a Learning Object Repository (LOR) stores only educational materials; and an e-Thesis Repository (TDR) stores only thesis and dissertations (Armbruster & Romary, 2010, Darby et al., 2009; Heery, 2009, Semple, 2006)). Question 3 (Q3) required the identification of the types of resources stored, i.e., books, papers, journal articles.

In section 2, where we sought to assess the use of international recommendations and MAPs by the DR, two questions were formulated: Q4) whether the repository adopts some sort of international recommendation – although not directly related to MAPs, its intention is to try to envision if the DRs community is open to the adoption of new recommendations and standards; and Q5) whether it adopts a MAP.

In section 3, we investigated which metadata schemas and elements are used by DRs. Therefore, we sought to determine: in Q6, which metadata schemas are adopted; in Q7-Q9, which DC, LOM and MODS elements, are adopted; and in Q10 which other schemas and elements are adopted. That way, we are able to draw an overview of what is being used and make relations, as well as achieve a parameter for future projects related to the definition of MAPs for DRs.

The questionnaire and its results may be accessed at the RepositóriUM, the Institutional Repository of the University of Minho (<http://repositorium.sdum.uminho.pt/>) by following the handle <http://hdl.handle.net/1822/35527>.

3. Results and Discussion

From 2,165 emails sent to the DRs' managers with a link to the questionnaire, 66 (3.1%) emails returned (wrong email address, not existent, et cetera). From the remaining (N= 2.165), 431 questionnaires were answered, corresponding to 19.9% of the total delivered.

The first question is about type of repository, 401 questions were answered and 30 were ignored. Of the total (n=401), 69 respondents (17.20%) identified their repository as being of more than one type. From these, 9 are indicated as OR and IR. We believe that, in this case, respondents might not have fully understood the differences between OR and IR. Therefore, we sought OpenDOAR in order to decide to which typology each of these 9 repositories should be assigned. After this exercise we verified that, from the total (n=401), the IR are prevalent (358, or 89.27%), followed by the TDR (52, or 13%), TR (36, or 9%), OR (25, or 6.23%) and LOR, (15, or 3.74%).

Four hundred and fourteen (n=414) DR managers answered Q3, while 17 left it blank. Scientific articles are identified as the most stored type of resource (350, or 84.54%), followed by books/chapters (320, or 77.29%) and theses and dissertations (318, or 76.81%). Respondents also informed about the storage of: datasets, media appearances, administrative and technical documents, blogging academics, curricula and other grey literature. Additionally, it was mentioned the use of metadata of journal of articles. Informal conversations with DRs managers at conferences and other events made it clear that some of them consider that a platform that only has metadata (and not contents) should not be considered a DR.

Comparing the types of repositories and the types of resources stored, it is clear that DRs are storing several types of resources, regardless of their pre-defined typology as answered in Q2 (Figure 1). Also, the results show that not all kinds of resources are subject to a quality control process such as peer review, which confirms Heery's claims (2009, pp. 13).

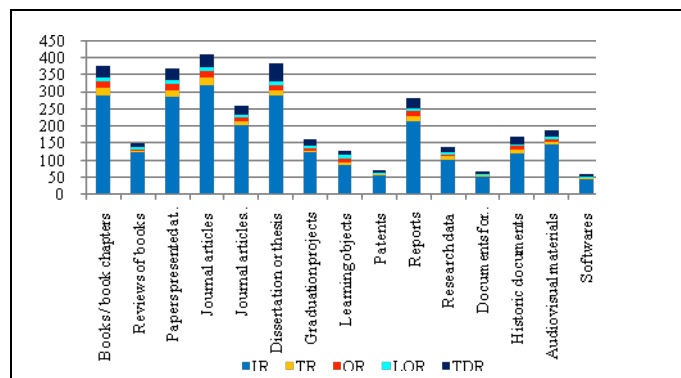


FIG. 1. Type of resources mostly stored by Digital Repositories

As to Q4, section 2, 376 questions were answered and 55 left blank. From total (n=376), some respondents claim to use DCMI recommendations (314, or 85.51%) and the OAI-PMH protocol (308, or 81.91%). A relatively low number of DR adopts SKOS (7, or 1.9%) and OWL (7, or 1.9%). It is noted, however, a greater number of those using RDF specifications. In accordance with results, the OAI-ORE standard has on IRs their biggest supporters (30, or 93.75%).

On the "other options", the respondents also quoted the Digital Repository Infrastructure Vision for European Research (DRIVER) recommendations. In the same field the respondents mentioned the use of other supporting documentation, not all classifiable as recommendations. These include specific APs, metadata schemas, data models, encoding/markup languages, file

formats, frameworks to create and use self-defined metadata formats, as follows: Guidelines SNRD Del Ministerio de Ciencia, Tecnología e Innovación productiva -Argentina; RIOXX, European Semantic Elements 3.4.1- ESE and European Data Model 5.2.6-EDM, JSIC-Eprints Metadata Model; EThOS UKETD-DC; VOA3R; XMetadiss, XMetaDissPlus; Open Language Archives Community-OLAC-DC; Component Meta Data Initiative-CMDI, MarcXml; Encoded Archival Description Document-EAD; TagSuite NLM DTD; NISO Z39.96-2012; JATS XML; Google scholar metadata tags; OpenAire; bibtext; schema.org; Digital Commons Metadata; MODS+ORE, Open Archives Initiative Static repository.

The answers to Q5 indicate that the number of repositories having defined APs is still very low. Overall, 342 questions were answered and 89 were ignored. From the total answered, two hundred and ninety-two (292, or 85.38%) respondents stated that their repositories do not adopt MAPs and 50 (14.62%) responded that they do. From these 50, 46 (13.5% of n=342) signaled "YES" (has MAP) and 4 (1.1% of n=342) signaled "Yes" and used the comment box to express their doubts as to what would be a MAP. Additionally, from the 46 affirmative answers, it was not possible to confirm the existence of a MAP for 25 (7.31% of n=342), even by following the URI that 6 of them provided; the existence of MAPs was confirmed only for 13 (3.8% of n=342) by using the URI they provided (Table 1). These results were obtained after we have analyzed each of the repositories on which there was an indication of the existence of MAPs and only the MAPs that fit NISO (2007) definition were taken into account. Table 1 presents the URIs of the 13 identified MAPs that are being used by these 13 DRs. It is worth mentioning that from the 89 that did not answer this question, 10 (11.2% of n=89) stated that they did not know what a MAP was. Summing these 10 with the above 4 in the same conditions, there was an overall of 14 respondents that claimed to not know the meaning of Application Profile. Although this number is very low (3.25% of n=431), it is reasonable to suppose that more respondents could have this doubt despite the definition was available just before the question.

TABLE 1. Application profiles used by Digital Repositories

REPOSITORY IDENTIFICATION	URI OF IDENTIFIED MAPs
Edinburgh ResearchArchive (ERA)	http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=The +EThOS+ UKETD_DC+application+profile
Kagoshima University Repository	http://www.nii.ac.jp/irp/en/archive/pdf/junii2_en_20090213.pdf
Rutgers University Community Repository	https://rucore.libraries.rutgers.edu/collab/reference.php?group=ALL&auth=ALL&type=ap&submit=Search
BibliotecaValenciana Digital	EDM 5.2.4 and EUROPEANA
ScienceCentral	http://www.e-sciencecentral.org/pub/pubinfo/
University of Oslo Open Res.Archive	https://www.cristin.no/openaccess/Dokumenter/Metadata_handbok_final.pdf
Biblioteca Digital de Castilla y León	http://www.digibis.com/software/digibib.html
BRAGE HihmHøgskoleniHedmark	http://brage.bibsys.no/xmlui/handle/11250/92963
UOC Repositori Institucional	http://openaccess.uoc.edu/webapps/o2/bitstream/10609/8055/6/GRISSET_metadadesUOC_2010_cat.pdf
REDICCES	http://www.redicces.org.sv/jspui/bitstream/10972/1763/1/guia_metadatos.pdf
Alaskas Digital Archives	https://scholarworks.alaska.edu/page/policy
DSpace at Rice University	https://digitalriceprojects.pbworks.com/w/page/89346902/Research%20Data%20Management%20Application%20Profile
Europe PubMed Central	http://dtd.nlm.nih.gov/2.0/xsd/archivearticle.xsd

Although we could not find similar studies, we found others that resemble in some way. Park & Tosaka (2010), for instance, obtained results that indicate a high percentage of MAPs usage within Digital Repositories + Digital Collections. Smith-Yoshimura & Cellentani (2007) found a low level of adoption of MAPs in digital libraries. None of these results can be directly compared to ours, once the objects are quite different. A study by Curado Malta & Baptista (2014) only found 10 MAPs specifically built for libraries and DRs and 31 for Learning Objects applications, that although not directly comparable to ours, corroborates its main finding: the low level of adoption of MAPs in the DRs community... Furthermore, both Park & Tosaka (2010) and Curado

Malta & Baptista (2014) report difficulties in accessing MAP related documentation, that in the case of Curado Malta & Baptista was partly solved by making direct contact with the MAP managers.

With regard to metadata schemas adopted (Q6, section 3), the prevalence is the Dublin Core Metadata Element Set (DCMES - reported by some respondents as simple DC) (269, or 83.80%), followed by Open Archives Initiative-Dublin Core (OAI-DC) (131, or 40.81%), Metadata Encoding and Transmission Standard (METS) (43, or 13.40%) and Machine-Readable Cataloguing (MARC) (39, or 12.15%), Metadata Object Description Schema (MODS) (36, or 11.21%), Electronic Thesis and Dissertation Metadata Standard (ETDMS) (30, or 9.35%), Learning Object Metadata (LOM) (13, or 4.05%), DSPACE intermediate metadata (DIM) (11, or 3.48%), Multimedia Content Description Interface (MPEG-21) (7, or 2.18%) and Academic Metadata Format (AMF) (2, or 0.62%). The Open Digital Rights Language (ODRL) and Metadata schemas for exchanging business cards (vCard) were not used by any of the DR (Figure 1). Here it is worth clarifying two aspects. The first is what was termed the DC Simplified and Qualified by the respondents. Many people still calls DC qualified (expression fallen into disuse within the DCMI community) to the set of DC elements plus its refinement elements (now all included in the DC Metadata Terms vocabulary - <http://www.dublincore.org/documents/dcmi-terms/>). It should be noted that the DSpace platform includes the so-called "DC Qualified" metadata elements, some of them not belonging to DC Terms and that were set as part of the development of this platform. The second aspect relates to the DC and OAI-DC: OAI-DC is the way the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) calls the 15 elements of the Dublin Core Metadata Element Set. Therefore, we combined these results considering that OAI-DC, DCMES and Simple DC are, in fact, referring to the same vocabulary/schema.

In the field "Others", the respondents also indicated the use of the following schemas: unofficial Croatian metadata scheme (based on CROSB1); XMetaDiss; Date Document Initiative (DDI 3.2); Directory Interchange Format (DIF); CIF core dictionary; MTD2-BR; hal.fr; Digital Item Declaration Language (DIDL); Component Metadata Infrastructure (CMDI); Darwin Core for the Virtual Herbarium collection; Text Encoding Initiative (TEI); Registry Interchange Format – Collections and Services (RIF-CS); Research Document Information Format (ReDIF). However, as in Q4, some answers do not really correspond to metadata schemas: Document Object Model (DOM), GNU Eprints, Collex.org; World Bank-specific taxonomies, and Google Scholar Metadata, and the already mentioned OLAC, ESE, EDM; ORE; JATS DTD.

These results show that: a) a great number of repositories store different types of resources (398, or 99.25%), which means that elements drawn from one or more metadata schemas could probably be used as a complement to DC, in order to enhance the description of those resources. Some of these repositories, however, only use DC; b) some repositories use metadata elements drawn from two or more schemas. In this case it could be advisable to define a MAP; c) the usage of LOM elements is more visible in IRs than in LORs, prevailing the use of DC in all of them.

The prevalence of the use of DC might be justified with the results of Q4 that show the data collection is based on the OAI-PMH protocol, which uses only DC by default. There are metadata schemas designed for specific and detailed descriptions, potentially enabling resources' "find ability" and more relevant and precise search results (Heery & Anderson, 2005, Vogel, 2014). Organizations such as DCMI and W3C offer recommendations for "mixing and matching" these elements into a coherent whole and in a machine-readable and interoperable way. By using different metadata schemas repositories' managers can optimize the information exchange between the various information services. In addition to MAPs, it is worth noting the recent W3C developments on the Shapes Constraint Language (SHACL), which is an RDF vocabulary to identify RDF graphs' "predicates and their associated cardinalities, data types and other constraints" (Knublauch, et al., 2015). A Draft version was recently published that contains use cases and requirements (Steyskal & Coyle, 2015)

The results of Q7 show that most of the 15 DC elements are highly used by DRs (Figure 2). In addition to the 15 elements, the respondents also indicated the use of the following DC Terms elements: alternative (1, or 0.32%), bibliographicCitation (1, or 0.32%), isPartOf (1, or 0.32%) and audience (1, or 0.32%). Respondents also informed about elements that are not part of DCTerms, that were added by DSpace [sic]: placeOfPublication (1, or 0.32%), root (1, or 0.32%), series (1, or 0.32%), number, edition, volume; level of audience; dc.contributor.author; dc.subject.other (1, or 0.32%); author contact (1, or 0.32%); editor contact (1, or 0.32%); date available (1, or 0.32%); date accessioned (1, or 0.32%); start page (1, or 0.32%); end page (1, or 0.32%); ispartofname (1, or 0.32%); ispartofnumber (1, or 0.32%); ispartoftitle (1, or 0.32%); ispartofvolume (1, or 0.32%), level of audience (1, or 0.32%); open access (1, or 0.32%), embargo (1, 0,32%). One respondent informed that he “incorporated other metadata elements in records for ETDs”. Another respondent extended DC in order to include information about “media of materials and number of pieces and NBN identifier”. This is an old practice that was already identified by Heery e Patel (2000) who have claimed that implementers use metadata schemas pragmatically and that this procedure in the past started with the use of MARC, when implementers introduced their own fields, instead of adopting the concept of “mixing and matching schemas”.

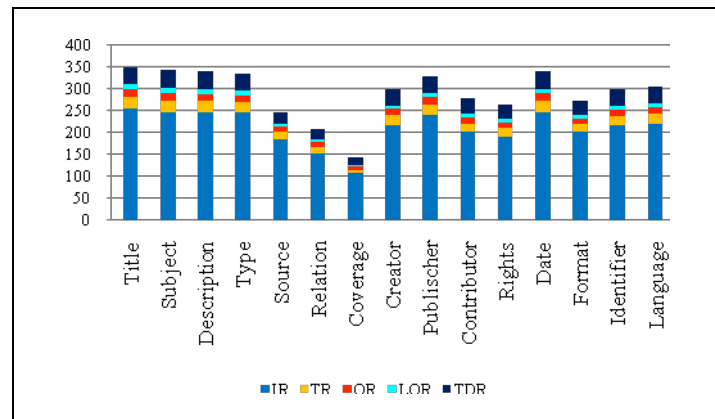


FIG. 2. DC elements used by Digital Repositories

As for LOM, the categories most frequently used were General and Educational (Figures 3 and 4). Some elements were used only by just one repository.

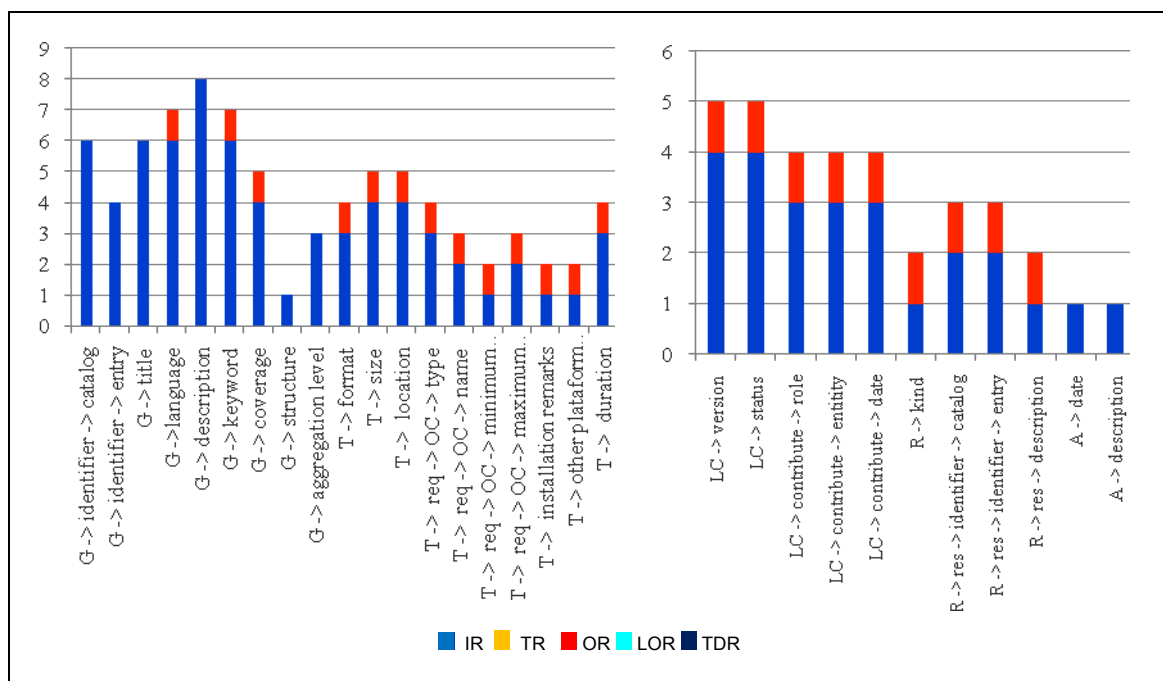


FIG. 3. LOM elements used by Digital Repositories (General, Technical, Life Cycle, Relations categories)

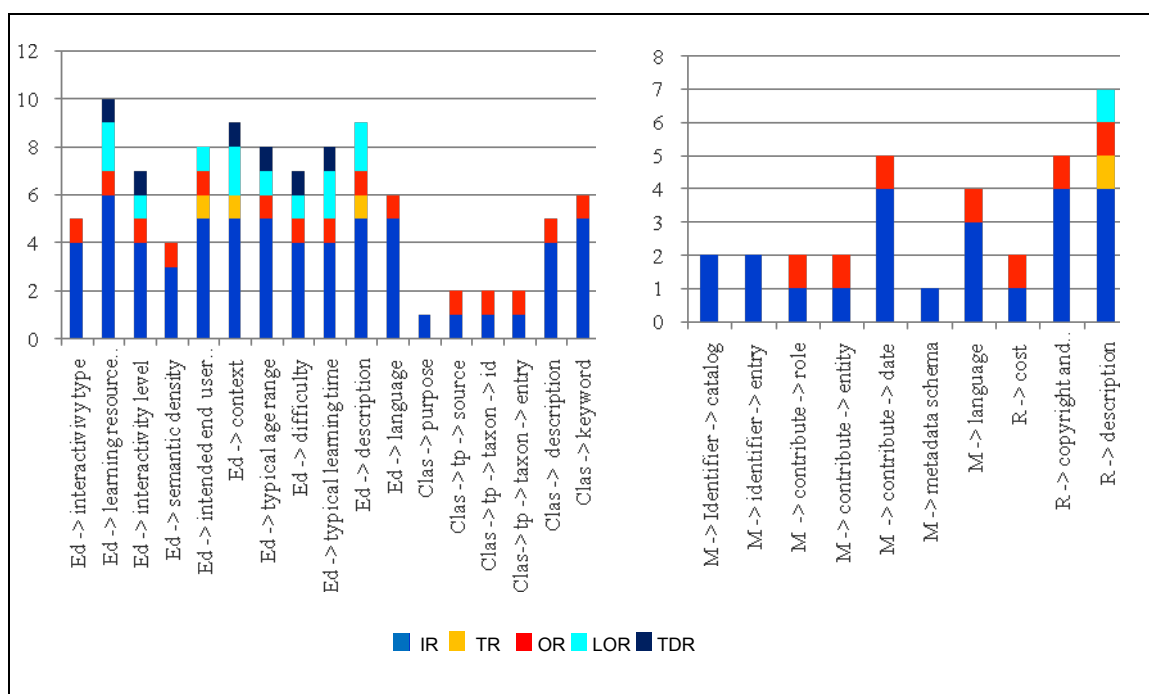


FIG. 4. LOM elements used by Digital Repositories (Educational, Classification, Meta-Metadata e Rights categories)

MODS elements were adopted by (29, or 6.72%) DRs. The IRs use more MODS elements than any other type of repository (Figures 5 and 6). This fact maybe related to its compatibility with MARC 21, which is widely used in the libraries' domains (Assumpção & da Costa, 2013). The fact that MODS was developed for the description of bibliographic resources, considering the

libraries domain ("Metadata Object Description Schema", n.d.), contributes for its adoption by information professionals.

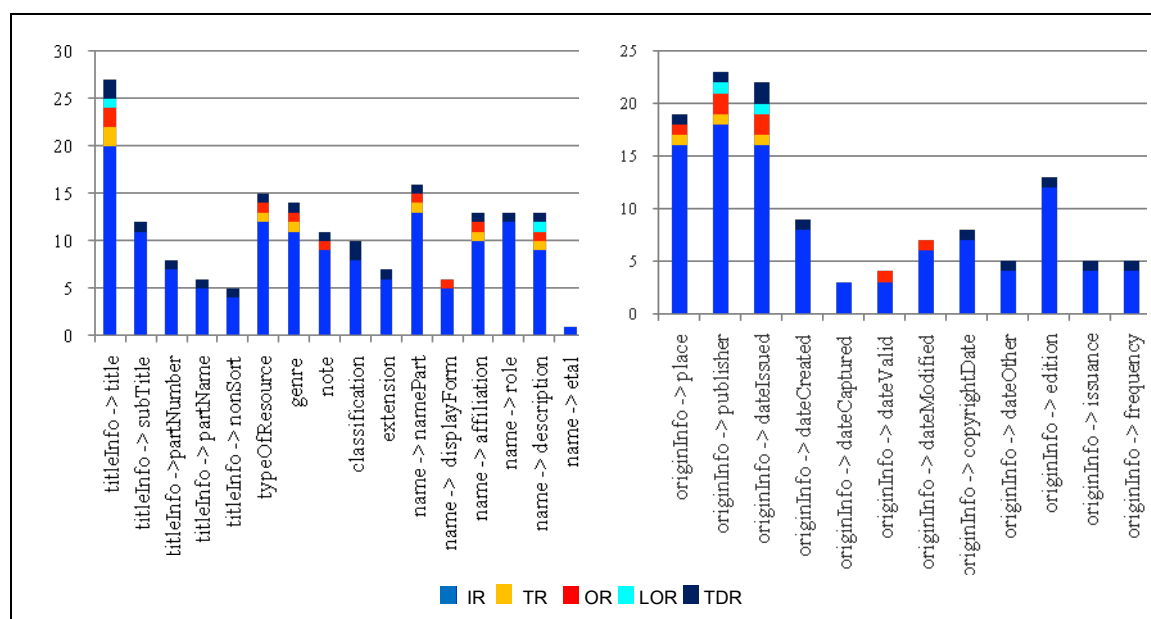
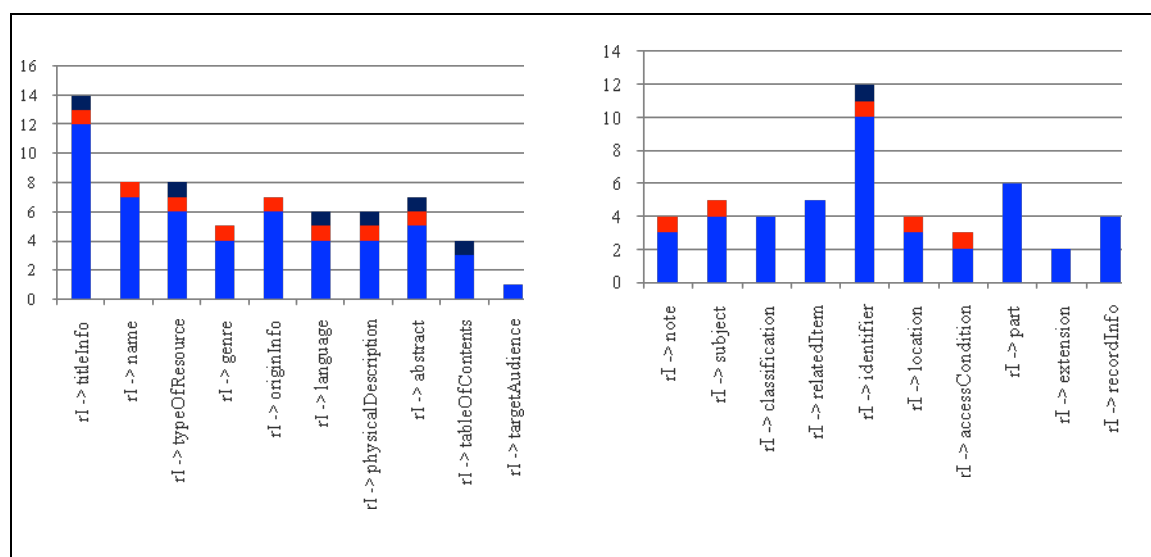
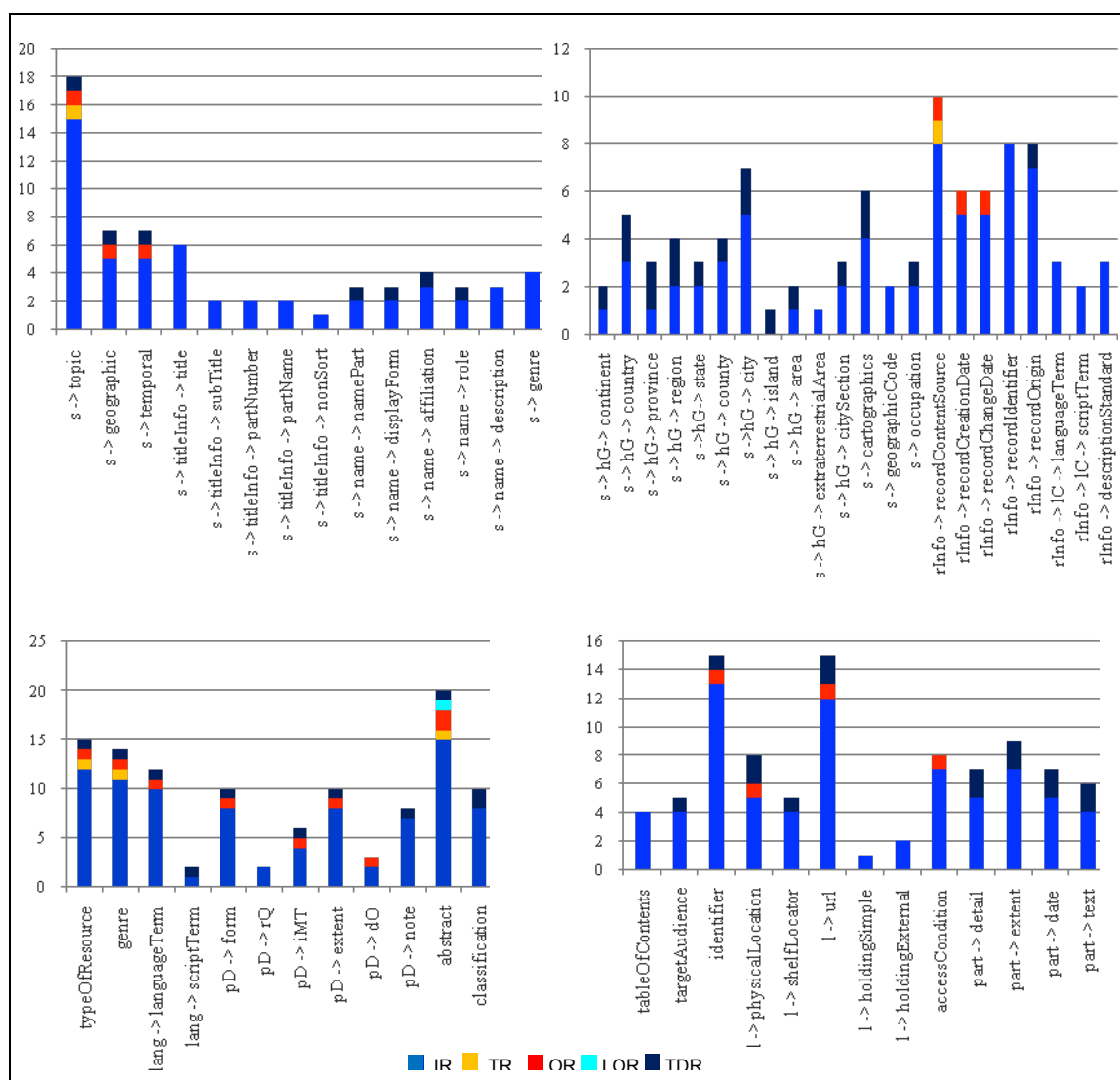


FIG. 5. MODS elements and subelements used by Digital Repository (titleInfo, typeOfResource, genre, note, classification, extension, name, originInfo Elements)





Legend: lang – language; pD – physicalDescription; l – location; rQ – reformattingQuality; iMT- internetMediaTypes; d-O, digitalOrigin; rI – relatedItem; s – subject; rInfo – recordInfo, IC – languageOfCataloging

FIG. 6. MODS elements and subelements used by Digital Repository (relatedItem, subject, recordInfo, typeOfResource, genre, language, physicalDescription, tableOfContents, targetAudience, identifier, location, accessCondition, part Elements).

Q10 is open: the respondents could inform about other schemas and elements being used and that were not previously mentioned in the questionnaire. The results are presented in Table 2.

TABLE 2. Other metadata schemas and elements used by Digital Repositories

Type of DR	Metadata Schema	Metadata Elements
IR / TDR	ETD-MS (NDLTD)	thesis.degree.level thesis.degree.name etd.degree.discipline etd.degree.grantor etd.degree.level etd.degree.name etd.thesis.degree title, director, advisor)
	Elements used to better capture resources by Google Scholar (citation_title citation_author

		citation_online_date citation_pdf_url
IR / TDR	https://rucore.libraries.rutgers.edu/rutgers-lib/30699/record/	rulib:descriptiveEvent->type rulib:descriptiveEvent->dateTime rulib:descriptiveEvent->detail rulib:descriptiveEvent->associatedEntity rulib:descriptiveEvent->associatedObjectand similar events in sourceMD, techMD, and rightsMD
IR / OR	elementsdevelopedinternally	utb.event.state utb.identifier.wok utb.faculty utb.identifier.scopus utb.source utb.identifier.obdid utb.identifier.rivid

The results presented in Table 2 are related to three different situations: a) Two DRs use metadata elements drawn from other schemas or created by them, but they do not have MAPs explicitly created; b) One DR has a MAP and makes it publicly available; and c) One respondent claimed his repository had a MAP, but it is not accessible.

Conclusion

The data collected shows that:

- the number of repositories that define APs, is very low, regardless of their typology, contrasting with DCMI recommendations that recommend the use of MAPs in order to optimize semantic interoperability. The lack of knowledge by managers about the advantages and the definition of APs might be one of the factors that inhibit its adoption;
- IR is the type of repository using a greater variety of metadata schemas and using them more. However, we realize that while others follow the trend of the IR, LOR and TDR do not exploit so much the metadata schemas that have been developed for their predominant resource types;
- Dublin Core Element Set is the most adopted metadata schema. Other schemas quite used are METS and MARC 21. This result may be justified by: a) the simplicity of DC and by the fact that it is the schema used by default by OAI-PMH; b) METS simplicity, extensibility and modularity; and 3) the history of MARC 21 in the information science discipline.

The five most used elements in a) DC: title, author, description, date and type; b) LOM: General -> description, General -> identifier -> catalog, General -> title, General -> language, Educational -> learning resource type; 3) MODS: titleInfo -> title, originInfo-> publisher, abstract , originInfo ->dateIssued and subject -> topic;

- The respondents show a lack of knowledge about MAPs and its adoption.

Limitations and future study

The main limitations of the study are:

- limited number of answers. Although we have achieved a considerable number of respondents (431 out of 2,165), many questionnaires were not completely answered (111, or 25.8%), and many questions were left blank. The questionnaire was quite dense and some questions, such as the ones related to MAPs, might be considered complex for some DR managers. The contributions of other agents that participate in DRs' management might have been useful although it is our belief that the MAP concept is not well disseminated in the DRs community.
- lack of knowledge by the respondents about some concepts touched on some questions, despite of the almost totality of questions have been explained as to their meaning. This is a situation that deserves repositories' specialists and managers

attention, once the lack of knowledge of some themes inhibits the progress of the actions that can strengthen and optimize the use of the open access through RDs, the semantic interoperability and the adoption of Linked Data guidelines.

Future studies could focus in identifying which metadata schemas and elements are being used by different resource types in DRs. In addition, future studies could include the usage of interactive tools as the wiki.

Acknowledgements

We thank Espírito Santo Federal University, Brazil; CAPES Foundation, Ministry of Education of Brazil for financial support to our research activities. Part of this work has been supported by FCT –Fundação para a Ciência e Tecnologia within the project scope UID/CEC/00319/2013.

References

- Armbruster, C., & Romary, L. (2010). Comparing repository types. Retrieved from <http://arxiv.org/ftp/arxiv/papers/1005/1005.0839.pdf>
- Assumpção, F. S., & da Costa, P. L. V. A. (2013). Metadata Authority Description Schema (MADS): uma alternativa à utilização do formato MARC 21 para dados de autoridade; Metadata Authority Description Schema (MADS): una alternativa al uso del formato MARC 21 para datos de autoridad. *Informação & Informação*, 18(1), 106–126.
- Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. Retrieved from <http://www.ecommons.cornell.edu/handle/1813/7895>
- Chan, L. M. (n.d.). Metadata Interoperability: A Study of Methodology. Retrieved July 15, 2015, from <http://www.white-clouds.com/iclc/cliej/cl19chan.htm>
- Clayphan, R., & Oldroyd, B. (2005). Using Dublin Core application profiles to manage diverse metadata dDevelopments. In *International Conference on Dublin Core and Metadata Applications* (pp.–23). Retrieved from <http://dcpapers.dublincore.org/index.php/pubs/article/view/800>
- Curado Malta, & Baptista, A. A. (2014). A panoramic view on metadata application profiles of the last decade. *International Journal of Metadata, Semantics and Ontologies*, 9(1), 58. <http://doi.org/10.1504/IJMSO.2014.059124>
- Darby, R. M., Jones, C. M., Gilbert, L. D., & Lambert, S. C. (2009). Increasing the Productivity of Interactions Between Subject and Institutional Repositories. *New Review of Information Networking*, 14(2), 117–135. <http://doi.org/10.1080/13614570903359381>
- DCMI Usage Board. (2009, March 2). DCMI Usage Board Review of Scholarly Works Application Profile. DCMI. Retrieved from <http://dublincore.org/usage/reviews/2009/swap/>
- Diamantopoulos, N., Sgouropoulou, C., Kastrantas, K., & Manouselis, N. (2011). Developing a Metadata Application Profile for Sharing Agricultural Scientific and Scholarly Research Resources. In *Metadata and Semantic Research* (pp. 453–466). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24731-6_45
- Guenther, R. (2000, August 6). DC-Library Application Profile (DC-LAP). Retrieved from <http://dublincore.org/documents/2001/08/08/library-application-profile/>
- Heery, R. (2009). Digital Repositories Roadmap Review: towards a vision for research and learning in 2013. Retrieved from <http://kennison.name/files/zopestore/uploads/libraries/documents/reproadmapreviewfinal.pdf>
- Heery, R., & Anderson, S. (2005). Digital repositories review. Retrieved from <http://opus.bath.ac.uk/23566/2/digital-repositories-review-2005.pdf>
- Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, 25, 27–31.

- Hillmann, D. I., & Phipps, J. (2007). Application profiles: exposing and enforcing metadata quality. Retrieved from <https://ecommons.library.cornell.edu/handle/1813/9371>
- Knublauch, H., TopQuadrant, Inc., Prud'hommeaux, E., & W3C/MIT. (2015, April 2). Shapes Constraint Language (SHACL). W3C Editor. Retrieved from <https://w3c.github.io/data-shapes/shacl/>
- Metadata Object Description Schema: MODS (Library of Congress). (n.d.). Retrieved July 27, 2015, from <http://www.loc.gov/standards/mods/>
- National Information Standards, N. (2004). Understanding metadata. *National Information Standards*, 20.
- National Information Standards Organization. (2007). *A framework of guidance for building good digital collections*. Bethesda: NISO.
- Nilsson, M., Baker, T., & Johnston, P. (2008, January 14). The Singapore Framework for Dublin Core Application Profiles. Retrieved from <http://dublincore.org/documents/singapore-framework/>
- Park, J.-R., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: schemata, selection criteria, and interoperability. *Inf. Technol. Libr*, 29(3), 104–116.
- RIOXX the RIOXX metadata profile and guidelines. (2014). Retrieved March 24, 2015, from <http://rioxx.net/v2-0-final/>
- Semple, N. (2006). Digital Repositories. Retrieved July 16, 2015, from <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/digital-repositories>
- Smith-Yoshimura, K., & Cellentani, D. (2007). *RLG Programs Descriptive Metadata Practices Survey Results: Data Supplement*. OCLC Programs and Research. Retrieved from <http://www.wip.oclc.org/content/dam/research/publications/library/2007/2007-04.pdf>
- Steyskal, S., & Coyle, K. (2015, July 16). SHACL Use Cases and Requirements. Retrieved July 22, 2015, from <https://w3c.github.io/data-shapes/data-shapes-ucr/>
- Vogel, D. M. (2014). Qualified Dublin Core and the Scholarly Works Application Profile: A Practical Comparison. Retrieved from <http://digitalcommons.unl.edu/libphilprac/1085/>



Metadata Praxis—Session 5

A DDC Visual Interface for Metadata Exploration

Xia Lin, Michael Khoo,
Jae-wook Ahn
Drexel University,
USA
{xlin, mjk326, ja626}
@drexel.edu

Ceri Binding,
Douglas Tudhope
University of South Wales,
UK
{ douglas.tudhope, ceri.binding}
@southwales.ac.uk

Hilary Jones,
Diana Massam
MIMAS, University of
Manchester, UK
{Hilary.Jones, Diana.Massam}
@manchester.ac.uk

Abstract

This paper presents a visualization interface for DDC-enriched metadata collections. Three sets of metadata from three different digital libraries were aggregated and re-indexed. Automatic analysis was performed to assign one or more DDC classes to each individual metadata record. A comprehensive search and exploratory interface was designed and implemented to include dashboard views, localized views, and universe views of DDC and the metadata collections. Finally, an experiment was conducted to test and compare how subjects interacted with different views for metadata search, exploratory and resource discovery.

Keywords: visualization interface; metadata exploration; digging into metadata; Dewey Decimal Classification; DDC; automatic classification; interface testing and evaluation.

1. The Digging Into Metadata Project

As one of the “Digging into Data” projects (Digging into Data Challenge, n.d.), the *Digging Into Metadata* project investigated innovative methods for metadata enhancement and reuse. The project was conducted among our three research groups in the last three years (January 2012 to December 2014). It addresses three crucial needs of enhancing metadata for finding, retrieving, and sharing digital resources: the need to aggregate metadata records in multiple digital libraries, the need to perform automatic analysis of metadata collections and use the results to enhance individual metadata records, and the need to create new interfaces to access digital resources through the enhanced metadata.

An assumption being tested in this project is that some knowledge organization systems (KOS) can be mapped automatically to a collection of metadata to enhance semantic connections among the metadata records. The test bed for the project is the mapping of Dewey Classification System (DDC) numbers to an aggregated set of metadata from three digital libraries: the National Science Digital Library (U.S.A.: <http://nsdl.org/>) (also including the Digital Library for Earth Systems Education, DLESE: <http://www.dlese.org/>); the Internet Public Library (USA: <http://www.ipl.org/>) (also including the Librarians’ Internet Index (LII); and Intute (U.K.: <http://www.intute.ac.uk/>). The IPL was founded in 1995 in the U.S. as an online reference service, and then began developing digital collections (Janes, 1998). In 2008, the IPL merged with the Librarians’ Internet Index (LII), and the IPL and LII metadata was crosswalked to Dublin Core and added to a Fedora database (Khoo & Hall, 2010). The NSDL is an NSF-funded federated multi-disciplinary STEM library, with a central Dublin Core metadata repository currently housed at <https://nsdl.oercommons.org/>. The library includes metadata from a number of individual domain-specific portals, or ‘Pathways’ (e.g. Zia, 2004; Bikson et al., 2011). As the NSDL Pathways were independent entities, the same resource could be cataloged in different ways by different Pathways. Finally, Intute was developed in the U.K. by a grass-root community dedicated to online educational resource discovery (Joyce, 2008; Williams, 2006). Much Intute metadata was inherited from a series of previous partners and educational consortiums in the U.K., and as a result, each Intute resource has both a Dublin Core record, and can also have

additional subject classification metadata stored in separate SQL tables, partly a legacy of previous specific subject catalogs to suit the needs of users of particular collections. Each of these digital libraries therefore had histories dating back to at least the early 2000s. Further, the metadata for each collection included the standard Dublin Core elements (*title*, *description*, *subject*, *identifier*, etc.), although due to the large number of contingencies in the histories of each library, each catalog also contained a range of qualified elements. The harvesting and normalization processes were therefore quite complicated (Khoo et al., forthcoming).

For this purpose, we have designed and tested a workflow pipeline that includes: (1) harvesting metadata records; (2) extracting metadata from designated fields in each record; (3) analyzing this metadata and generating weighted key terms that represent 'aboutness'; (4) using these weighted key terms to generate one or more DDC numbers that can then be added back to the records concerned; and (5) using the new DDC numbers in each records to build tools that allow users to search and browse multiple DDC classes at the same time. The design and discussion of this pipeline has been reported in (Binding, et al., 2013, Khoo, et al., forthcoming).

The mapping of DDC to metadata records has also been reported in (Khoo, et al., 2012). For this process, two major components were developed: MASH (Metadata Analysis, Sharing, & Harvesting), and DISTIL (Document Indexing and Semantic Tagging Interface for Libraries). The MASH component includes the process of (1) cleaning metadata records harvested from multiple digital libraries, (2) extracting nouns from selected metadata elements of each record, (3) calculating Term Frequency (TF) scores after applying known language processing procedures such as tokenization, stop word removal, and stemming, and (4) ranking the noun list using a TF-score based weight schema. As a result, MASH produces a ranked list of nouns that can be sent to DISTIL for bulk analysis.

The goal of the DISTIL component is to generate automatically one or more DDC class numbers for each metadata record, which can then be used to support searching and browsing. DISTIL follows a document classification approach with two main phases. The first phase attempts to match a weighted combination of the key terms in a metadata record against the entry vocabulary of DDC. This results in many matches both across different DDC hierarchies and at different levels within a given hierarchy. The second phase takes account of matches within hierarchies, aggregating lower level matches to broader parents. Depending on the configuration, outliers without any ancestor or descendant matches can be discarded. Essentially, DISTIL determines an overall degree of match between two sets of records: a metadata record, and DDC class headings, including DDC Relative Index headings. It then generates an output for each metadata record with the top N DDC numbers assigned to that record.

This paper reports the third part of the project on designing and developing new interfaces that will take advantages of the enhanced metadata (the metadata records with DDC numbers assigned) for resource discovery across multiple heterogeneous digital collections.

2. The Interface for DDC-enriched Metadata

Having the metadata records with DDC numbers automatically assigned has the potential to facilitate searching, browsing, and resource discovery. To fulfill the potential, specialized user interfaces need to be created (Slavic, 2006). When discussing desirable interface functions to support the use of classification systems for searching and browsing, Slavic emphasized that the advantages of using a hierarchical/facet classification for browsing and retrieval depend on the strength of the interface -- "The power of the interface is in supporting visualisation that will 'convert' what is potentially a user-unfriendly indexing language based on symbols, to a subject presentation that is easy to understand, search and navigate."

Creating visual exploratory interfaces that integrate classifications, subject terms, and search results is therefore a major goal of *the Digging into Metadata* project. In particular, we envision that:

- The interface should take full advantage of DDC classification structures to create “views” to guide the user. DDC has well-built hierarchical structures and associative class relationships. The structures may be used to create global views of the collections and localized views of queries and search results.
- The interface should utilize the new associations among metadata records as a result of assigning multiple DDC classes to metadata records. When metadata records are associated with DDC classes, new associative relationships are formed through the DDC-metadata relationships, which can also be converted into new metadata-metadata relationships. Both relationships might be used to guide user’s searching and browsing activities, including querying and filtering.
- The interface should support user’s interaction with both DDC class structures and search results. The use of DDC structures should help, rather than hinder, the user’s interactions. The user does not need to be familiar with DDC structures, and the user should have choices of what they want to “see” and when to “see” or use the DDC knowledge structures.

These three requirements have become our design principles for creating the new interfaces. While they look simple and straight forward, the implementation has proved to be quite challenging.

2.1. The Interface and its Components

To build the interface, a solid indexing and searching platform was needed. We chose the open-source package, Solr (<http://lucene.apache.org/solr>), for this purpose. Three different data types are indexed in Solr for the project:

- Metadata: titles, descriptions, subject descriptors, and URLs of the digital resources
- DDC: DDC class, division, and section numbers and the class labels delivered from DISTIL
- Digging Statistics: context analysis results such as term frequencies and scores, DDC numbers co-occurrence frequencies, etc.

Currently, the Solr platform provides access to about 79,500 metadata records from IPL, Intute, and NSDL. These are the records that have been analyzed and for each of them one or more DDC classification numbers are assigned.

The front-end was built as a web application with html, JavaScript, and visualization tools such as D3.js (<http://d3js.org>) and Sigma.js (<http://sigmajs.org>). Figure 1 shows a sample display of the interface. As shown in this example, the interface contains three major parts, as well as the top bar where the user may enter a search query. On the left is the area for DDC tree display; on the right, the bottom part is the display area for retrieval results and the top part is a tabbed area for three different visual widget displays, which are described next.

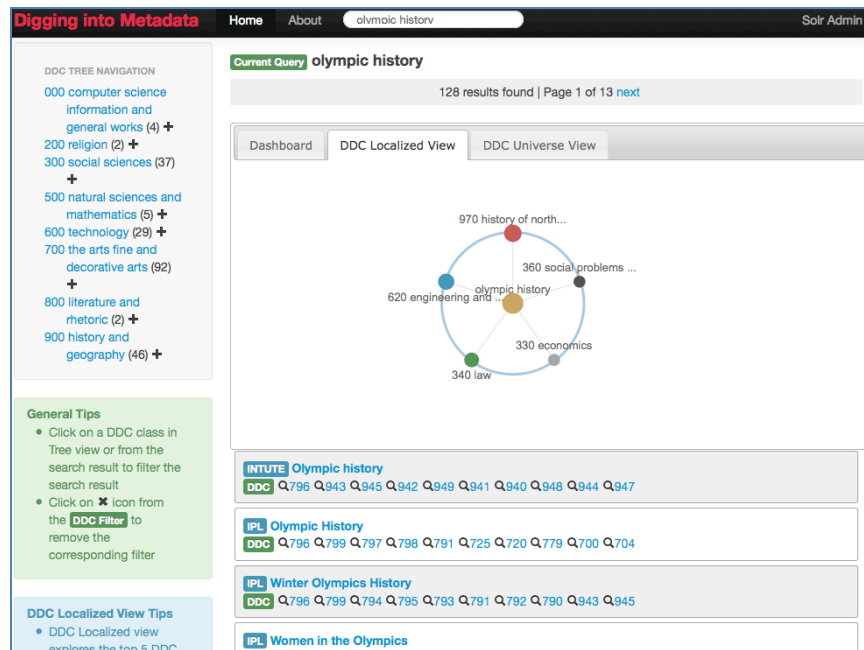


FIG. 1. A sample display of the digging user interface

2.2. Three Interactive Widget Displays

Using tabs is an effective way to provide alternative views of information within a limited display space. In figure 1, the tab selected is the “DDC localized view” which shows on a circle the 5 closest DDC classes related to the query. The circle format was chosen for its simplicity, interactivity, and easy repetition. In this example, the query “Olympic history” is most closely related to {DDC970, History of North America; 620, Engineering & allied operations; 340, Law; 330, Economics; & 360, Social problems & social services}. This list serves as the most succinct interpretation of issues related to the query “Olympic history.” Displaying it on a circle is much easier to read and interact with. The circle effectively extends the query to a “query ring” with which the user can interact to refine his or her searches. For example, the user may click on a DDC number to bring up another “ring” with a new set of relevant DDC numbers, and he or she may choose a DDC number on the rings to add to the query to narrow down the search results. To a large extent, the usefulness of the DDC “query ring” will depend on the accuracy of the automatic DDC class assignment created by the DISTIL process. It will provide similar functions like those “synonym rings” described in Zeng (2006).

The other two tabs also provide unique functions for the user to interact with both search results and related DDC classes. Figure 2 shows the Dashboard display for the same search query “Olympic history.” The display shows the top 10 DDC numbers occurring in the retrieved results, arranged by their occurrence frequency. The user can, for example, click on “725, public structures” to retrieve 55 documents which is the result of search query “Olympic history AND DDC:725.”

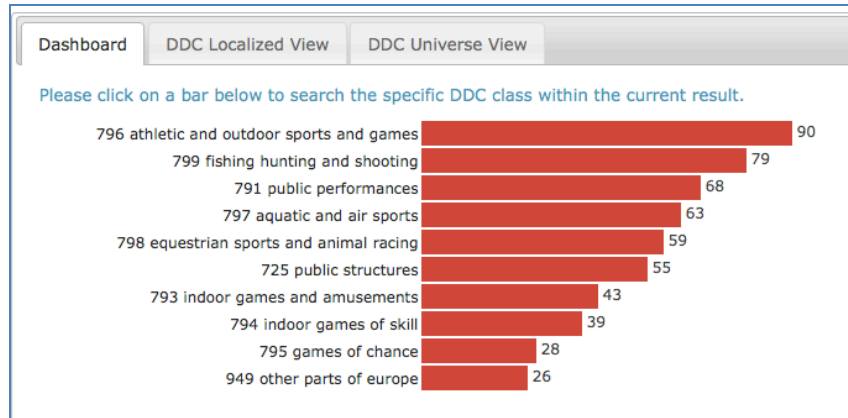


FIG. 2. The Dashboard display for the search query “Olympic history”. It lists the top 10 DDC classes of the retrieved results by their occurrence frequencies.

The third tab, the “DDC Universe View”, provides more elaborate functions to explore the network of metadata records (figure 3). While the DDC localized view is a bottom-up approach to exploring the collection – the user starts from a query and moves iteratively towards the target – the DDC Universe View is a top-down approach. It starts with showing the entire DDC universe of the underlying digital collections as a large graph in which DDC classes are represented as nodes and documents are depicted as links. Through the DDC Universe View, the user can zoom in and out dynamically, pan horizontally or vertically, or jump to a specific location of the network by a given DDC class. The user may (1) explore how a given DDC class is connected to other DDC classes in this collection, (2) learn what are the major clusters of the whole collection and what are the main classes within each cluster, and (3) locate seemingly unrelated but interesting new relationships by following the links. These activities help *serendipitous* discovery of new connections and benefit the exploratory nature of search. In this example, when the user zooms in to the first DDC class in the search result Dashboard display (DDC 796), he discovers that DDC 307.3 (“structure; Abandoned buildings”) is one of the closest DDC nodes to DDC796. This indicates that the topic on “Abandon buildings” and “athletic and outdoor sports and games” is one of the common themes in this collection. The user may choose these two nodes to find the resources.

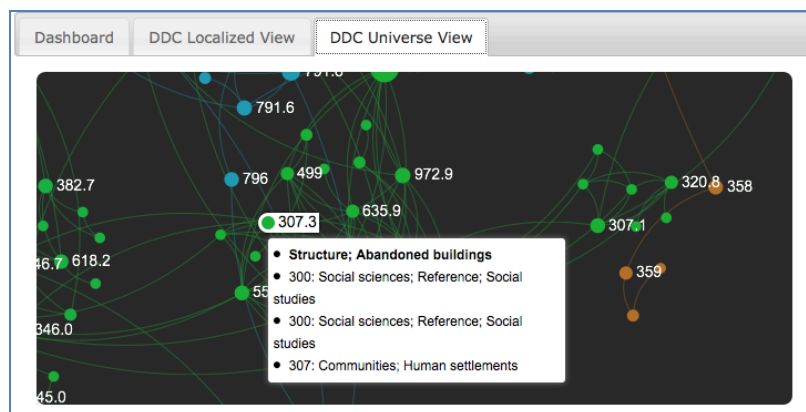


FIG. 3. DDC Universe View for the search query “Olympic history”. When the user zooms in, he discovers that DDC 307.3 (“structure; Abandoned buildings”) is closely related to DDC796 (“athletic and outdoor sports and games”) in this metadata collection.

3. Testing and evaluating the interface

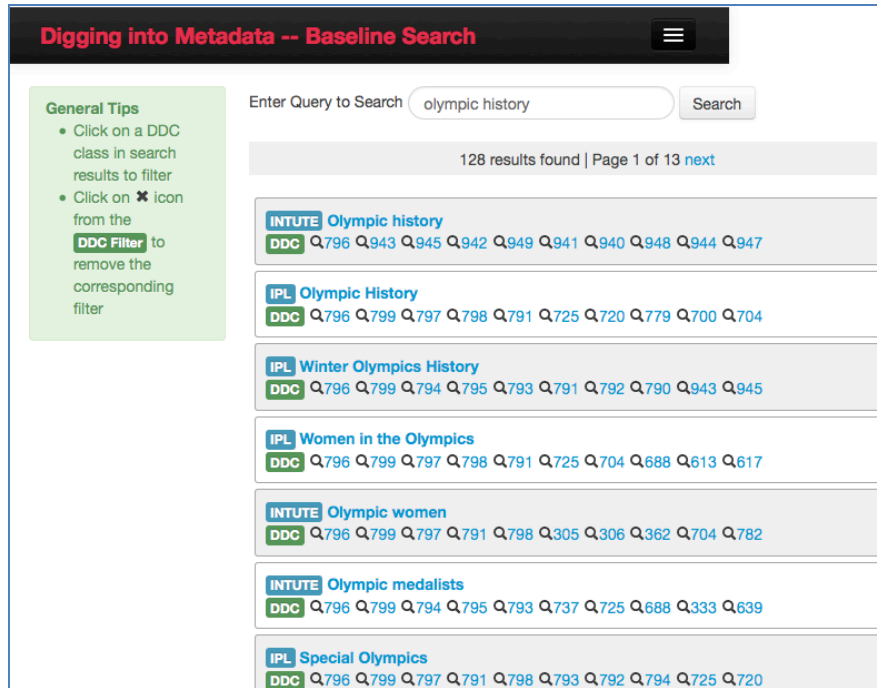
The interface was designed to serve several purposes. First, the interface will allow the user to access multiple metadata collections created with different standards or metadata structures, once these collections have been re-indexed. Second, the visual interactive functions of the interface will support a new way of exploring metadata collections through DDC distributions and their relationships to metadata records. Third, the interface will be a good testing platform to study how DDC may be applied to support searching, browsing and exploration of metadata collections. In this section, we reported our first effort in testing how users interact with the interface in an experimental setting.

To isolate the interactive functions we planned to test, we first separated the interface into three different implementations, each with a unique way of using the DDC classes for searching, browsing and exploration. In this experiment, the first interface is a simple search interface that returns search results with associated DDC numbers (the search interface; Figure 4(a)). The second adds a clickable DDC hierarchical tree to the search interface (the tree interface; Figure 4(b)). The third interface shows an interactive visual map of the search results and relevant DDC numbers (the visual interface – Figure 5).

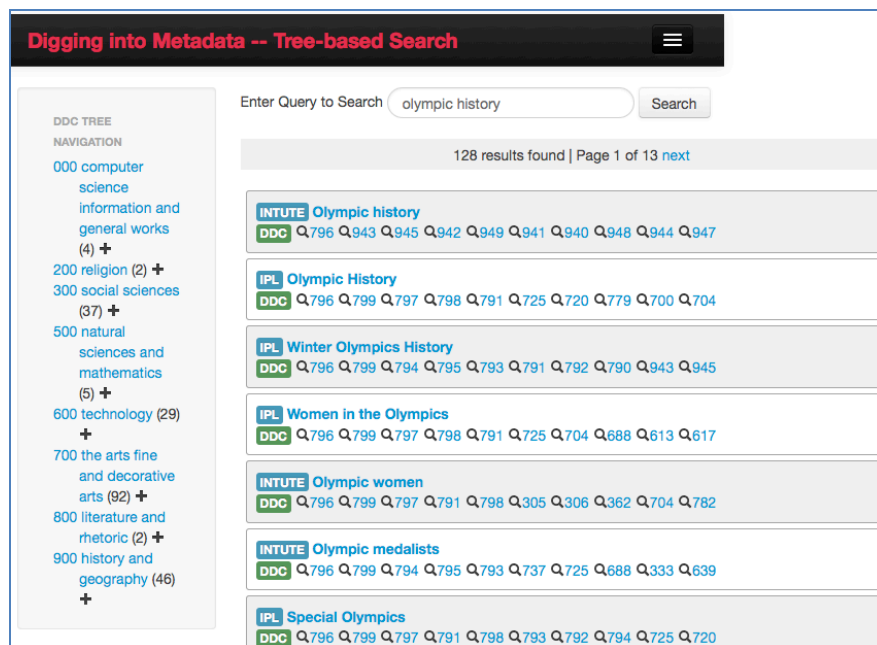
After getting the Institute Review Board (IRB) approval, we recruited 30 subjects, mostly undergraduate students (22 males and 8 females), for the experiment. They were paid \$10 each for the experiment that lasted for about an hour. Each of them first completed a pre-questionnaire and watched a short video that introduced the three different styles of the interface. They were then asked to complete one search task with each of the three interfaces. The same three search tasks (see Appendix), and the three interfaces, were rotationally assigned to the subjects to avoid any order impact or bias. After completing a search task, each subject completed an interface-specific post-questionnaire that asks questions such as how easy to use the search interface, how useful the DDC specific functions, whether they have a positive experience with the interface, and how satisfied are they with the search results, etc.

As the first step of data analysis, we focused on comparing the three different interfaces and how the subjects interacted with DDC classes shown on the interfaces. Two main results are reported here.

The first concerns the general impression of the interfaces. The results indicate that the subjects understood how to use DDC to filter or narrow down search results on all three of the interfaces. As shown in Figure 6, the subjects favored the tree interface consistently across the four categories: Easy of use, Usefulness, Positive experience, and Satisfaction with the search results. The differences, however, are small and not significant. While the search interface is perceived more easy to use than the visual interface, the visual interface seems to have achieved more satisfied results and is perceived more useful than the search interface. In some of the verbal comments that the subjects made, they also confirmed that the visual interface both most interested and most confusing to them, but they managed to use it nevertheless.



(a) Experimental interface 1 -- The search interface



(b) Experimental interface 2 -- The tree interface.

FIG. 4. The experimental interface 1 and 2.

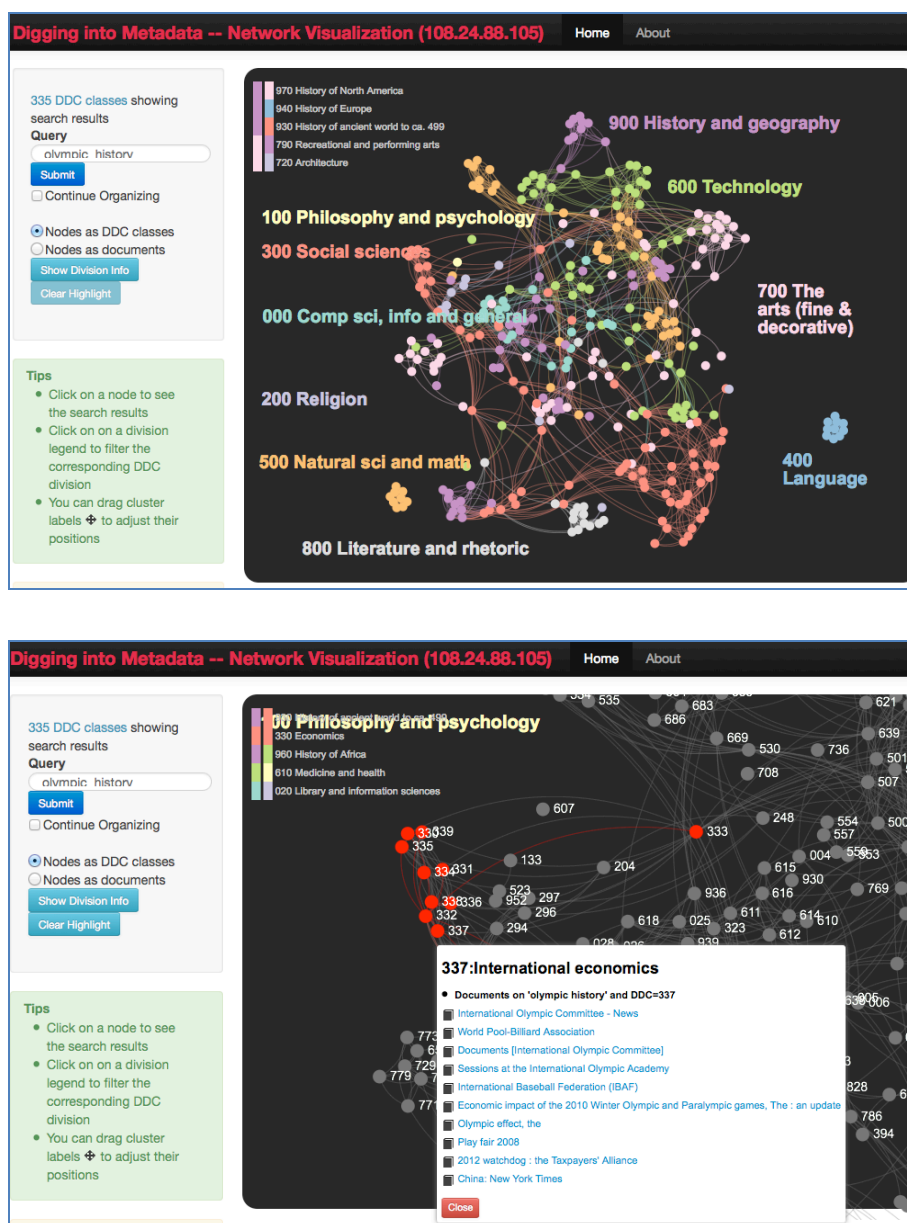


FIG 5. The experimental interface 3. The top one is the initial visual view of the interface 3 used in the experiment for the query “Olympic history”. The bottom one is the zooming view when the user clicked on the DDC label “330 Economics” and then DDC node “337, International economics”.

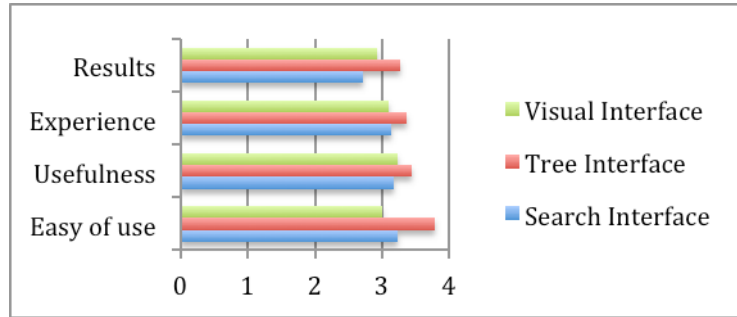


FIG. 6. Subjects' responses to post-task questionnaires. On a scale of 1 to 5, users were asked to rank each interface for its easy of use, usefulness, positive experience, and satisfaction of results. The tree interface is consistently better in all the four categories.

Table 1: DDC classes chosen for the three different interfaces

Search Tasks	Interface	DDC divisions chosen to explore (each subject could choose more than one DDC class)
Nuclear Testing	Search	3 (subjects) opted 620 (Engineering) 3 opted 621 (Applied physics) 3 opted 623 (Military & nautical engineering)
	Tree	5 opted 621 (Applied physics) 4 opted 628 (Sanitary engineering) 3 opted 623 (Military & nautical engineering)
	Visual	2 opted 572 (Biochemistry) (all other classes were selected by only one subject)
Water Cycles	Search	5 opted 551 (Geology, hydrology, meteorology) 4 opted 628 (Sanitary engineering) 3 opted 333 (Economics of land & energy)
	Tree	5 opted 550 (Earth sciences & geology) 5 opted 551 (Geology, hydrology, meteorology) 4 opted 628 (Sanitary engineering) 4 opted 333 (Economics of land & energy) 3 opted 577 (Ecology) 3 opted 621 (Applied physics)
	Visual	4 opted 628 (Sanitary engineering) 3 opted 620 (Engineering) 2 opted 631 (Specific techniques; apparatus, equipment, materials)
Hurricane Katrina	Search	5 opted 363 (Other social problems & services) 5 opted 551 (Geology, hydrology, meteorology) 4 opted 973 (United States) 3 opted 979 (Great Basin & Pacific Slope region of United States)
	Tree	2 opted 970 (History of North America) 2 opted 973 (United States) 2 opted 979 (Great Basin & Pacific Slope region of United States) 2 opted 620 (engineering)
	Visual	3 opted 970 (History of North America) 3 opted 973 (United States)

		3 opted 976 (South central United States)
		3 opted 324 (The political process)

The second result relates to how the three interfaces make the subjects “see” different DDC classes for the same search. With each interface, the subjects started by entering their own queries for the search task. Based on what they saw on the interface, they could choose one or more DDC classes to add to the query, or select DDC classes to explore related resources. Table 1 shows the DDC classes chosen by the subjects for each search task (Only those classes chosen by multiple subjects were shown in the table). Clearly, most of the DDC classes are relevant to the topics. The interfaces did have significant impacts on what the subjects perceived as relevant DDC classes to the query. The classes identified by the search and tree interfaces are significantly overlapped; however, the visual interface seems to lead to unique and diverse classes. Different subjects tended to see different relevant DDC classes with the visual interface. The tree interface also helps the subjects focused on the same branch of the DDC hierarchy (such as 621, 623, 628, and 550, 551, etc.).

4. Discussions & Conclusions

As the overall goal of our project, we successfully integrated three sets of metadata from different digital libraries, created a set of tools and procedures to automatically assign one or more DDC classes to individual metadata records, and established a new indexing service to provide access to the enhanced metadata collection with richer semantic connections. We believe that a new interface is still needed in order to make the best use of the new metadata collection.

Building DDC-based interactive interfaces have been reported in a number of cases (Pollitt & Tinker, 2000; Chowdhury and Chowdhury, 2004). There are also various research projects on taxonomy-based interfaces where hierarchical structures and categories of terms or concepts are utilized for searching and browsing (Khoo, Wang & Chaudhry, 2012). Another example is the metadata interface for enhanced metadata records with additional terms generated by a Topic Modeling algorithm (Hagedorn, Chapman, and Newman, 2007). The authors in particular discussed the benefits and limitations of using automated classification techniques to enrich metadata for searching and browsing. Building on similar ideas, our design goal is to integrate multiple views of DDC hierarchical structures, query-based contextual structures, and classification-based semantic structures for the purpose of interactive searching, exploration and discovery.

We have built a prototype interface to demonstrate the feasibility of such integration (available for testing at: <http://mcd.ischool.drexel.edu/ddcvisual>). Testing and evaluation of such interfaces, however, remains a challenge. For an interface for metadata exploration, there are issues of metadata integration and indexing, content representation and organization, and interactions and usability, to name just a few. All these issues have significant impacts on how well the interfaces could be used by users for their intended purposes. In the experiment reported here, we attempted to isolate some of the issues and focus on how users perceive DDC classes presented on the interfaces and how they used DDC classes for searching and exploration. Initial findings indicate that the subjects understand the values of DDC and found it useful for searching and exploration. They liked to interact with the DDC classes, and use them to filter the search results. The results also show that how DDC classes are presented on the interfaces will make a major difference.

Each of the three interfaces used in the experiment has some advantages and disadvantages. The search interface can quickly lead the user to see DDC classes most relevant to the user's query. The classification codes provide additional semantic links that the user might be able to follow to find relevant items. But in general the classification codes may not increase the precision for searching, as commented by a subject, “Once I was in a detailed topic I found it hard to return to look through broad DDC codes” (subject 9). The DDC tree interface, as another subject remarked, “is easy to use but it did not help much for this topic.” (subject 30). The data

showed that the visual graphs help the subjects see different DDC classes, but it is not clear that what the subjects saw was new insights that might not be seen in other interfaces. The visual interface “was most interesting but I feel like it was a bit hard to find the information that I wanted” (subject 21). Other comments indicate the visual interface was “very confusing” and with too much information, “the graph was easy to understand, but a lot of things were unrelated to the search” (subject 23).

While we are inspired by the subjects’ favorable impressions of the interfaces, it is also clear that the interfaces have not yet optimized for making the best use of DDC structures in an interactive and visual setting. In the future, we plan to conduct more experiments to understand how subjects interact with the DDC classes. A new experiment will be run for more specific exploratory tasks. Additional experimental modules will be implemented to log user’s interactions with the interfaces. We hope that the detailed log analysis will help us understand further how significant DDC plays in completion of the exploratory tasks and what additional benefits that automatic DDC classification will bring to the metadata collections.

Acknowledgements

The funding support from IMLS and JISC to the “*Digging into Metadata*” project was gratefully acknowledged. Thanks also go to OCLC for permitting the use of DDC for this research.

References

- Bikson, T., Kalra, N., Galway, L., & Agnew, G. (2011). Steps Toward a Formative Evaluation of NSDL. RAND Technical Report. http://www.rand.org/content/dam/rand/pubs/technical_reports/2011/RAND_TR998.pdf.
- Binding, C., Tudhope, D., Ahn, J-W., Khoo, M., Lin, X., Massam, D., & Jones, H. (2013). Digging Into Metadata. 12th European Networked Knowledge Organization Systems (NKOS) Workshop at the TPD Conference, Valletta, Malta, Thursday 26th September 2013.
- Chowdhury, S. and Chowdhury, G. (2004) Using DDC to create a visual knowledge map as an aid to online information retrieval. In: 8th International ISKO Conference: Knowledge organization and the Global Information Society, 2004-07-13 - 2004-07-16, London. Avail from: <http://strathprints.strath.ac.uk/2624/1/strathprints002624.pdf>
- Digging into Data Challenge. (n.d.). Retrieved from <http://diggingintodata.org/>.
- Janes, J. (1998). The Internet Public Library: An Intellectual History. *Library Hi Tech*, 16(2), 55-68.
- Joyce, A., Wickham, J., Cross, P., & Stephens, C. (2008). Intute integration. *Ariadne* 55. <http://www.ariadne.ac.uk/issue55/joyce-et-al>
- Khoo, C.; Wang, Z.; & Chaudhry, A.S. (2012). Task-based navigation of a taxonomy interface to a digital repository. *Information Research*, 17(4), 2012. Available at: www.informationr.net/ir/17-4/paper547.html
- Khoo, M., Ahn, J. W., Binding, C., Jones, H., Lin, X., Massam, D., & Tudhope, D. (forthcoming). Augmenting Dublin Core Digital Library Metadata with Dewey Decimal Classification. Paper accepted in *The Journal of Documentation*.
- Khoo, M., & Hall, C. (2010). Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. 10th ACM/IEEE Joint Conference on Digital Libraries, Brisbane, Australia, June 21-25, 2010, pp. 361-36.
- Khoo, M., Tudhope, D., Binding, C., Abels, E., Lin, X., & Massam, D. (2012). Towards Digital Repository Interoperability: The Document Indexing and Semantic Tagging Interface for Libraries (DISTIL). *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*), 2012, Vol.7489, pp.439-444.
- Hagedorn, K., Chapman, S.; & Newman, D. (2007). Enhancing search and browse using automated clustering of subject metadata. *D-Lib Magazine*, 13(7/8). Available at: <http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html>
- Pollitt, A. S.; Tinker, A. J. (2000), "Enhanced view-based searching through the decomposition of Dewey Decimal Classification Codes", *Proceedings of the Sixth international conference of the International Society for Knowledge Organization*, 10-13 July 2000, Toronto, Canada. Eds. C. Beghtol, L. Howarth and N. J. Williamson. Würzburg: Ergon Verlag, 2000. (*Advances in Knowledge Organization* 7). 288-294.

- Slavic, A. (2006). Interface to classification: some objectives and options. (UDC paper)
<http://arizona.openrepository.com/arizona/handle/10150/105459> .
- Williams, C. (2006). Intute: The New Best of the Web. *Ariadne* 48. <http://www.ariadne.ac.uk/issue48/williams>.
- Zeng, M. L. (2008). Knowledge organization systems. *Knowledge Organization*, Vol. 35, No. 2-3: 160-182.
- Zia, L. (2005). The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. *D-Lib Magazine* 11(3). <http://www.dlib.org/dlib/march05/zia/03zia.html>.

Appendix: The search tasks used in the experiment:

1. Please find best Web resources that a high school student should read when working on a paper for nuclear testing sites and its impact to the environments. What DDC classes would be useful for this topic?
2. You have been asked to prepare a class project on the water cycle, and to identify some of the current environmental, social, political, and other issues associated with different stages of the water cycle. Please identify relevant web resources and DDC classes.
3. Hurricane Katrina was one of the largest storms to make landfall in the United States, and the costliest in terms of damage to New Orleans and other places. Your project is to collect information for writing a timeline for Hurricane Katrina. The timeline should not just focus on the storm itself, but also look at such issues as the history of New Orleans, the social and political issues that were raised after the storm, the reconstruction, how the storm has been remembered, how the storm has affected peoples' lives today, and so on.

Leveraging SKOS to Trace the Overhaul of the STW Thesaurus for Economics

Joachim Neubert
ZBW – Leibniz Information
Centre for Economics,
Germany
j.neubert@zbw.eu

Abstract

“What’s new?” and “What has changed?” are questions users of Knowledge Organization Systems (KOS), such as thesauri or classifications, ask when a new version is published. Much more so, when a thesaurus existing since the 1990s has been completely revised, subject area for subject area. After four intermediately published versions in as many consecutive years, STW Thesaurus for Economics¹ has been re-launched recently in version 9.0. In total, 777 descriptors have been added; 1,052 (of about 6,000) have been deprecated and in their vast majority merged into others. More subtle changes include modified preferred labels, or merges and splits of existing concepts. We here describe how these changes were tracked, making use of the published SKOS (Miles & Bechhofer, 2009) files of the versions, loading them into named graphs of a SPARQL endpoint and executing queries on them. An ontology supporting version and delta description and query formulation is introduced. High-level visualizations of aggregated change data and drill-downs to the actual concepts are presented. We finish with an outlook to the *skos-history* project², which generalizes and extends the methodology to different knowledge organization systems.

Keywords: KOS; thesaurus; versioning; version history; Linked Open Data; Semantic Web; SPARQL; named graphs; service description

1. Use cases for change tracking

Vocabularies published on the web – particularly vocabularies shared under an open license like the Open Database License³ used by STW – can be downloaded without notification of the publisher and may be in use in multiple places and scenarios. Only some of them are known to the publisher. So there is no way to know for a maintainer which of the changes made to a vocabulary may or may not break things down the line. Handling changes quietly within an organization, as it was a widespread practice for a long time, isn’t an option any more.

Several use cases for change tracing have been identified within the *skos-history* project:⁴

1. support for human indexers for adapting their subject indexing practice to the new version – the classical use case
2. support for re-indexing large sets of documents, in an automatic, semi-automatic or manual fashion – vocabulary changes may require dealing with already indexed documents retrospectively
3. support for the maintenance of vocabulary mappings – new mapping targets may have occurred or already mapped concepts may have been deleted or deprecated

¹ <http://zbw.eu/stw>

² <https://github.com/jneubert/skos-history>

³ <http://opendatacommons.org/licenses/odbl/>

⁴ <https://github.com/jneubert/skos-history/wiki/List-of-Use-Cases>

4. support for the maintenance of derived vocabularies (e.g., a subset covering a special interest field, for the use within an independent organization)
5. support for vocabulary-based automated indexing applications
6. support for search applications

Most of these use cases involve applications for which a machine readable input format is highly desired. A standard case, which can be handled automatically, is the replacement of changed preferred labels or notations for display purposes. Similarly, for obsolete descriptors, which have been merged completely into others, the indexing of documents can be switched automatically – while the update of a mapping to another vocabulary may require intellectual verification. Other types of changes, particularly a split of concepts, may require a complete review of already indexed documents.

Due to the possibly large efforts required for a migration to the latest version, at any point in time multiple versions of a vocabulary will be in use concurrently by different institutions. One goal of the set of practices described below is to enable users of a vocabulary to calculate the impacts a version upgrade will have in their particular scenario.

Section 2 of this paper outlines the basic approach of STW to vocabulary versioning. The method to track changes based on Linked Data is introduced and discussed in section 3. Based on this method, section 4 presents reports for tracking different types of individual changes, mostly focused on the first use case described above. Section 5 demonstrates the use of visualizations of aggregated data to understand high-level changes of the whole vocabulary. Section 6 introduces a detailed history of single concepts. Section 7 provides an outlook to future work.

2. Basic vocabulary versioning approach

Maintenance of STW is done within a custom application. During rework of larger parts of the vocabulary it may be in an inconsistent state. When such parts – such as e.g. “Money and financial markets” or “Information and communication”, which might span multiple sub-thesauri – were finished, a new version was published, bearing a version number (marked up as “owl:versionInfo”) and a version date (marked up as “dct:issued”). The URIs of the concepts stay stable (Hillmann, Sutton, Phipps, & Laundry, 2006) – however, the web pages for the concepts, which include RDFa semantic markup, are created and published for each version anew, and their URLs bear version numbers and language tags⁵. The rdf/xml and turtle expressions of a concept are versioned too. (Neubert, 2009) All files of previous versions remain accessible without limitations. As we observed that users often save the webpage address instead of the persistent URI as link, since 2010 the URI part “latest” instead of a version number has been serving as a default and redundant “symbolic link” to the latest version. All web pages of previous versions carry transparent “water-marks” to indicate their outdated status.

Once introduced, concepts are never deleted. Instead, obsolete concepts are stripped of all semantic relations, and are marked with a property “owl:deprecated true”. A textual hint such as “Deprecated (last used in version 8.04)” is added as a “skos:historyNote” (in HTML together with a link to the according page). When applying, a “dct:isReplacedBy” property links to a still existing concept into which the deprecated concept was merged.⁶

Since its start on the web in 2009, STW has published lists of changes in the form of plaintext files. Additionally, as the RDFa-enhanced web pages and the bulk SKOS downloads of every published version have been kept available, users had a chance to look up changes and compare versions of concepts manually. Unfortunately, the plaintext change lists were not linked to the actual concept pages. Furthermore, there was no way of filtering nor aggregating the information – let alone accessing it for any kind of machine processing.

⁵ e.g., <http://zbw.eu/stw/version/8.14/descriptor/11716-2/about.en.html> is the English version 8.14 page about the concept <http://zbw.eu/stw/descriptor/11716-2> “Infrastructure”.

⁶ e.g., <http://zbw.eu/stw/descriptor/12257-3>

This clearly was not well-suited to the large-scale changes going on within the vocabulary. One option would have been extending the custom maintenance application to log actions and produce more expressive change reports. Yet, an application-specific hard-coded solution would have made it difficult to experiment with different forms of change reports. And whatever the outcome, it would have served STW only.

3. Using SKOS files for version comparisons

Instead of logging change transactions as they occur during the maintenance process, we decided to compare the final SKOS files of differing versions. This means that the methodology described here should be applicable not only by producers, but also by consumers of vocabularies or interested third-parties, without the need of out-of-band knowledge buried in whatever internal vocabulary maintenance system.

The approach is founded on an RDF database of vocabulary versions and computed version deltas (as described below), which can be flexibly evaluated by SPARQL queries. These can take advantage of the quite regular and predictable structure of SKOS files (as opposed to arbitrary ontologies). The database, referenced as “version store” in the remainder of the paper, is based on RDF named graphs⁷ and created by the following steps:

1. load every version into a named graph
2. compute the delta between two versions and add it as two separate named graphs of insertions and deletions
3. add metadata describing and linking versions and deltas in a separate version history graph

For step 1, every triple/quad store which can deal properly with named graphs should suit – for STW we used Fuseki from the Apache Jena project. An additional experiment was conducted successfully with Sesame.

Step 2 can be approached by simply diff-ing version files in ntriples format by means of the operating system and splitting the result into insertions and deletions, or by creating the graphs directly in the version store by executing SPARQL update queries which ask for the triples of one version graph MINUS the triples of the other version graph. Neither method works well with blank nodes – they are perceived as deleted and inserted completely. However, their use should not be essential for any SKOS vocabulary, and if occurring (as in STW, where a few complex “use instead” notes were expressed using blank nodes), it makes most sense to filter them out.

⁷ <http://patterns.dataincubator.org/book/named-graphs.html>

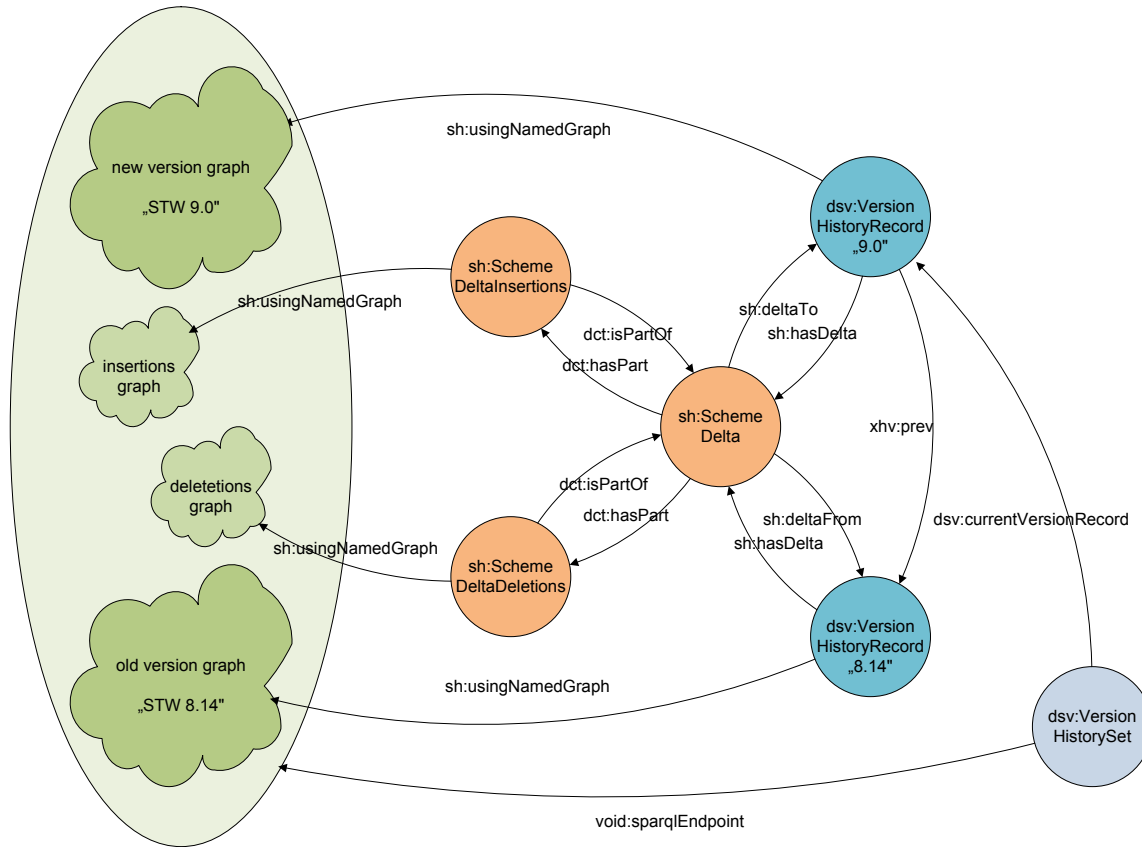


FIG. 1. Example data of the STW version history graph

For step 3, illustrated in FIG. 1, the emerging Dataset Versioning⁸ ontology (dsv) was used for describing the versions, furthermore the SPARQL 1.1 Service Description⁹ ontology (sd) for the named graphs and the newly developed skos-history¹⁰ ontology (sh) for the deltas and additional plumbing, supplemented by Dublin Core elements (dc) and terms (dct), the Vocabulary of Interlinked Datasets (void) and XHTML (xhv).

The basic idea of Dataset Versioning (derived from the ISO 25964 data model) is providing a “dsv:VersionHistoryRecord” (VHR) identified by a “dc:identifier” plus an optional “dc:date” for each version. Each VHR links to the single “dsv:VersionHistorySet” (these links are omitted in FIG.1), which in turn points back to exactly one VHR through a “dsv:currentVersionRecord” property. (De Smedt, Vrang, & Papantoniou, 2015) Pairs of VHR are connected via a “sh:SchemeDelta”, the chain of consecutive VHR is represented by “xhv:prev” links. The distinction between insertions and deletions parts, which link to the respective parts and named graphs, is implemented by different RDF classes.

For reference and re-use, all three steps for creating the version store are packaged in a publicly available bash script.¹¹ The version history set of STW is discoverable via a fix URI¹², as suggested in (ISO TC46/SC9/WG8 & Isaac, 2012). The embedded RDFa of the respective web page hints to the SPARQL endpoint containing the version store, which in turn includes the version history graph and the (default) service graph describing its overall structure.

⁸ <http://purl.org/iso25964/DataSet/Versioning#>

⁹ <http://www.w3.org/ns/sparql-service-description#>

¹⁰ <http://purl.org/skos-history/>

¹¹ https://github.com/jneubert/skos-history/blob/master/bin/load_versions.sh. The repository version referenced in this paper is tagged STW_9.0

¹² <http://zbw.eu/stw/version>

The method and metadata structure described here can be applied to every set of versions of a SKOS vocabulary – provided the versions are available as separate files and bear some kind of identifier. However, it is not well suited when vocabulary changes are published as a stream of update events, such as the subject updates published by the Library of Congress Linked Data Service as ATOM feed¹³. In case of LCSH it should be possible, however, to fall back to time-stamped download files of the whole data set.

4. Tracking version changes in change reports

The following change reports have been developed based on change categories proposed in (Pessala et al., 2011) and subsequently enhanced within the skos-history project.¹⁴ They operate on the version store and the metadata described in chapter 3. The queries used to generate the reports which are discussed in this chapter are publicly accessible.¹⁵ In a “SPARQL Lab” environment (Neubert, 2014), they can be loaded from GitHub and executed, but also inspected and modified by users.

Since the execution of the queries may take more than ten seconds, the results are cached as machine-readable JSON files (for which “raw data” download links are offered, too). The YASR¹⁶ library for Javascript supports formatted display of raw SPARQL results in browsers (Rietveld & Hoekstra, 2015). It provides additional user-friendly functionality, in particular the merge of concept URIs with their respective labels, presented as clickable links to the concepts (as shown in FIG. 2), furthermore paging for large result sets, and a quick search / filtering of the resulting tables.

4.1 Added and deprecated concepts

Added descriptors: This is the most basic information a report on vocabulary changes has to deliver. It can be obtained by asking for inserted concepts (identified by the occurrence of a “skos:prefLabel” triple in the insertions graph), for which no triples exist in the old version graph. The following SPARQL query can be executed against the public STW version store¹⁷:

```
# Identify concepts inserted with a certain version
#
SELECT distinct ?concept ?prefLabel
WHERE {
  # query the version history graph to get a delta and via that the relevant graphs
  GRAPH <http://zbw.eu/stw/version> {
    ?delta a sh:SchemeDelta ;
    sh:deltaFrom/dc:identifier "8.14" ;
    sh:deltaTo/dc:identifier "9.0" ;
    sh:deltaFrom/sh:usingNamedGraph/sd:name ?oldVersionGraph ;
    dct:hasPart ?insertions .
    ?insertions a sh:SchemeDeltaInsertions ;
    sh:usingNamedGraph/sd:name ?insertionsGraph .
  }
  # for each inserted concept, a newly inserted prefLabel must exist ...
  GRAPH ?insertionsGraph {
    ?concept skos:prefLabel ?prefLabel
  }
  # ... and the concept must not exist in the old version
  FILTER NOT EXISTS {
    GRAPH ?oldVersionGraph {
      ?concept ?p []
    }
  }
}
```

¹³ <http://id.loc.gov/techcenter/>

¹⁴ <https://github.com/jneubert/skos-history/wiki/List-of-Change-Categories>

¹⁵ <https://github.com/jneubert/skos-history/tree/master/sparql/stw>

¹⁶ <http://yasr.yasgui.org/>

¹⁷ <http://zbw.eu/beta/sparql/stwv/query>

This (simplified) query gives a list of all added concepts with their preferred labels in all languages. For use on the STW web site, the query is extended in several ways: The language is restricted to that of the current user interface, the concepts are restricted to descriptors (excluding subject categories – see below), and an additional column is added to provide information about the subject area to which the descriptor was added. Furthermore, the extended queries allow external parametrization via VALUES clauses and provide reasonable defaults (e.g.: compare the latest and the penultimate version of a vocabulary).

To this end, the overall structure of STW is exploited: Besides the actual descriptors with their poly-hierarchical broader/narrower relationships, it provides a mono-hierarchical system of subject categories (“concept groups” in terms of ISO 25964), which forms the sub-thesauri of STW and bear notations.¹⁸ Each descriptor is attached to one or more subject categories. For the majority of the change reports, the second level of this category system – e.g., “V.13 Labour” or “B.07 Marketing” – proved instrumental for breaking down the wide field of economics into about 80 meaningful subject areas.

Deprecated descriptors (with replacements): In a similar way, the deprecated descriptors are retrieved from the version store. This produces a table, which should be helpful for human indexers, but might also be used by scripts to update an already indexed database of documents.

secondLevelCategory	deprecatedConcept	replacedByConcept
V.03 Macroeconomics	Asset accumulation	Saving incentives
V.03 Macroeconomics	Consumption statistics	Household survey
V.03 Macroeconomics	Household expenditure	Private consumption
V.03 Macroeconomics	Intertemporal income distribution	Intergenerational mobility
V.03 Macroeconomics	Macroeconomic effect	Impact assessment

FIG. 2. Change report: Deprecated descriptors (extract)

Added subject categories / Deprecated subject categories (with replacements): Similar to the reports for descriptors, these track changes of STW’s category system. On a more global level, these reports expose where new fields of knowledge have emerged, or where on the contrary subdivisions are no longer regarded as necessary. For example, the subject category “W.19 Computer Software and Services Industries” (in version 8.04: “Data Processing”) was renamed to “W.19 ICT industry” and extended with further sub-ordinated categories, namely “W.19.3 Broadcasting Industry”, “W.19.4 Telecommunications” and “W.19.5 Information Services”. These new subject categories cluster already existing descriptors scattered over the category system before as well as newly introduced ones complementing the newly formed field of knowledge.

4.2 Label changes

Changed preferred labels: Since SKOS requires at most one preferred label per concept per language, we can safely identify cases where this label has changed. In this report, we offer links to the old and the new version of the descriptor, by constructing a version-specific URL to the corresponding web pages. This allows users to directly compare these pages.

Since for STW subject categories the preferred labels are created by prepending the label itself with the notation of the category, the changed preferred labels report for subject categories reflect

¹⁸ In SKOS, STW descriptors and subject categories are represented as subtypes of skos:Concept, namely zbwest:Descriptor and zbwest:Thsys (zbwest: <http://zbw.eu/namespaces/zbw-extensions/>)

also changes in the notation. This can reveal far-reaching changes in the STW category system, in the hierarchy or even in the assignment of partial category trees to sub-thesauri.

Added labels / Deleted labels: The former report shows all inserted preferred and alternate labels for descriptors (with the concept itself and its preferred label), the latter the deleted labels. Since labels do not carry an own identity and only the lexical values can be tracked, even minor changes in spelling (e.g., from “Advertising Industry” to “Advertising industry”) show up independently in the deletions and in the insertions list.

4.3 Hierarchical relations

Added narrower relationships / Added broader relationships: Changes in the descriptor hierarchy may be relevant in particular for newly inserted narrower concepts, so if the concepts to which the relationships were inserted are new, it is marked in these reports, too. Prior intermediate concepts in the hierarchy, which had been removed, are indicated also.

4.4 Other types of merges and splits

Splits: Labels moved to new descriptors: When a label is attached to another concept in a new version, this can be regarded as a hint that possibly the scope of the originating descriptor has changed. Particularly when the concept to which the label is attached is inserted with the same version, this may reveal a split of concepts (“Confidence interval”, for example, has been moved from “Estimation theory” to the newly introduced concept “Interval estimation”). This report, together with the Added-narrower one, can be taken as a basis for intellectual review of already indexed documents, which may or may not match the newly introduced narrower concept.

Merges: Labels moved from deprecated and split-up descriptors: This report covers the opposite situation, and particularly the case where the labels from a deprecated concept now are attached to other concepts than the one it was replaced by (e.g., the label “Royalties” was moved from the deprecated concept “Right of use” to “Charges”, while “Right of use” otherwise was merged into “Industrial property rights”). This can be taken as a hint that it may not be advisable to automatically re-index documents with the “replacedBy” concept without prior intellectual review.

Especially the two latest reports document a shift in the meaning of the remaining concepts, while their URIs have stayed the same. This may seem improper from a purely ontological point of view. However: In a real life environment defined by limited resources and cumbersome legacy library systems, nobody would want to take the additional effort to change already existing descriptors and re-index large amounts of documents without inescapable necessity.

Of course, the amount of necessary (re-)indexing is a factor which already has been taken into account up-front. Descriptors which have been used only a few times over the years are natural candidates for deprecation, while a re-indexing of thousands or ten-thousands of documents would not be considered with levity. Within ZBW, the holdings and the number of documents indexed with a particular descriptor are known, and the re-indexing can take place in parallel to the preparation of a new STW version. External parties using the thesaurus for indexing have to catch up afterwards. To provide information about the changes in an easily comprehensible way, to allow the estimation of subsequent efforts in local systems unknown to the editors of STW, is a primary goal of the approach described here.

5. Visualizing change with aggregated data

The change reports, as outlined above, allow tracking every single change from a certain class of changes. However, they are too detailed to facilitate insight into the development of the vocabulary as a whole. Yet, SPARQL 1.1 provides the means to query the version store for aggregated data. This allows the creation of statistics and charts which give a high-level overview over the changes to STW, particularly when aggregated over the complete amount of changes from version 8.06 to 9.0.

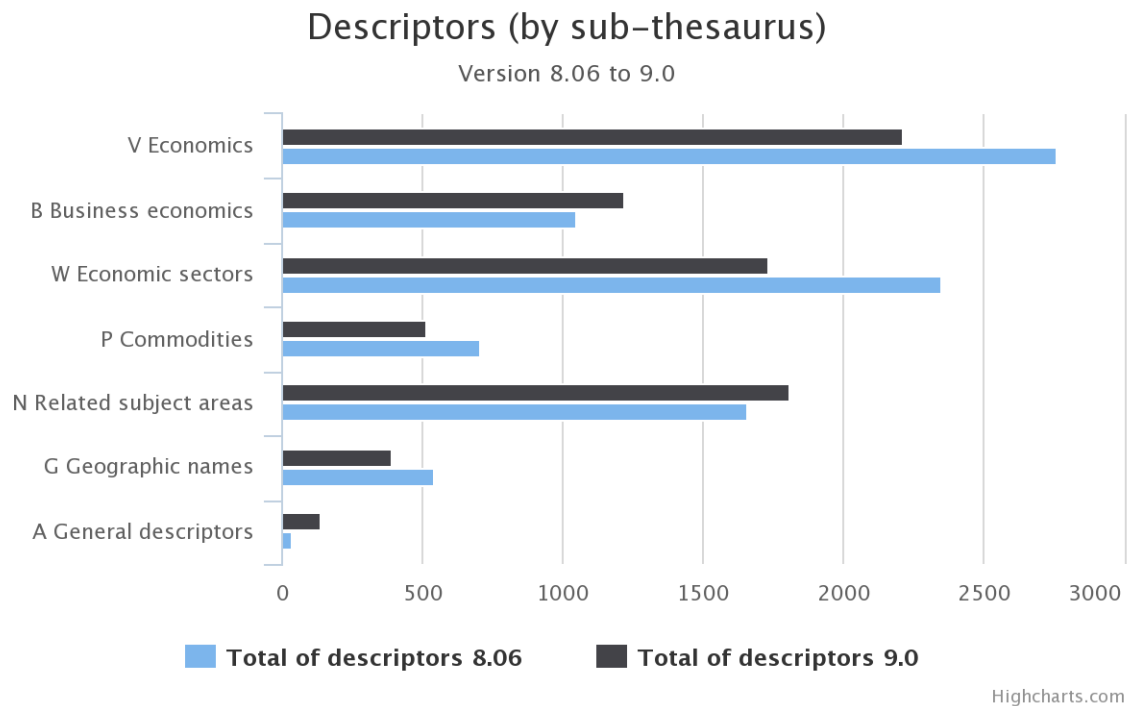


FIG. 3 Total number of descriptors

When we compare the total number of descriptors of these versions by sub-thesaurus (FIG. 3), we can see that the “Business economics” sub-thesaurus has been extended. The interactive graphics¹⁹ allow drilling down and discovering that the number of descriptors has grown particularly for the subject categories “Management and business organization”, “Logistics” and “Marketing”. The “Economic sectors” and “Commodities” sub-thesauri both have decreased counts of active descriptors. This is true for the general “Economics” sub-thesaurus, too. However, a more detailed analysis, facilitated by the “Changed preferred labels/notations” report on subject categories, reveals that this is partly caused by the movement of the whole field of mathematical and statistical methods from “Economics” to “Related subject areas”. On the whole, the branches of STW look more balanced after the overhaul.

While the overview charts give the net amount of descriptors, a series of more detailed charts (FIG. 4 and FIG. 5) shows the number of additions and deprecations within a certain part. These graphics support drill-downs, too.

In the example from the “Business economics” sub-thesaurus we can easily see that specifically in the fields of “Accounting” and “Corporate tax management” the deprecated/merged descriptors by far outnumber the added ones, and that a relatively small field such as “Operations research” has extended its coverage considerably.

The change graphics not only provide a high level overview. They work at the same time as a navigation tool, which allows focusing on the most interesting fields of change, and drill down into the change reports for added or deprecated concepts, by a passed-in search filter restricted to a particular 2nd level subject category. As the change reports link to the concepts themselves, this allows investigations up to the finest details.

¹⁹ linked from <http://zbw.eu/stw/relaunch>

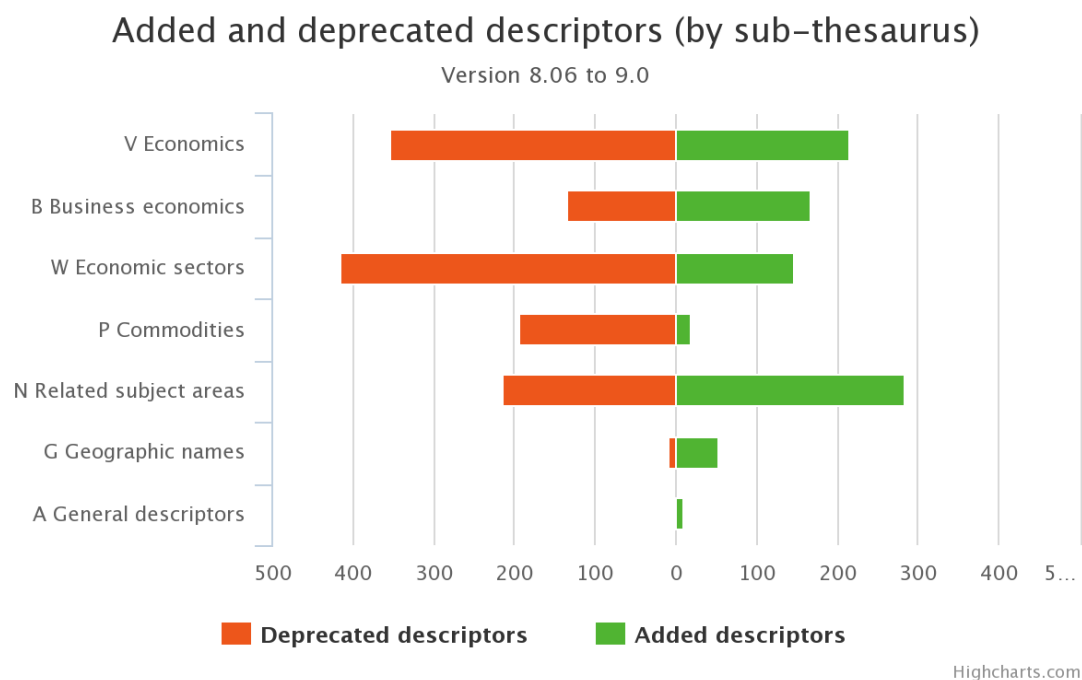


FIG. 4 Number of added and deprecated descriptors by sub-thesaurus, with drill-downs

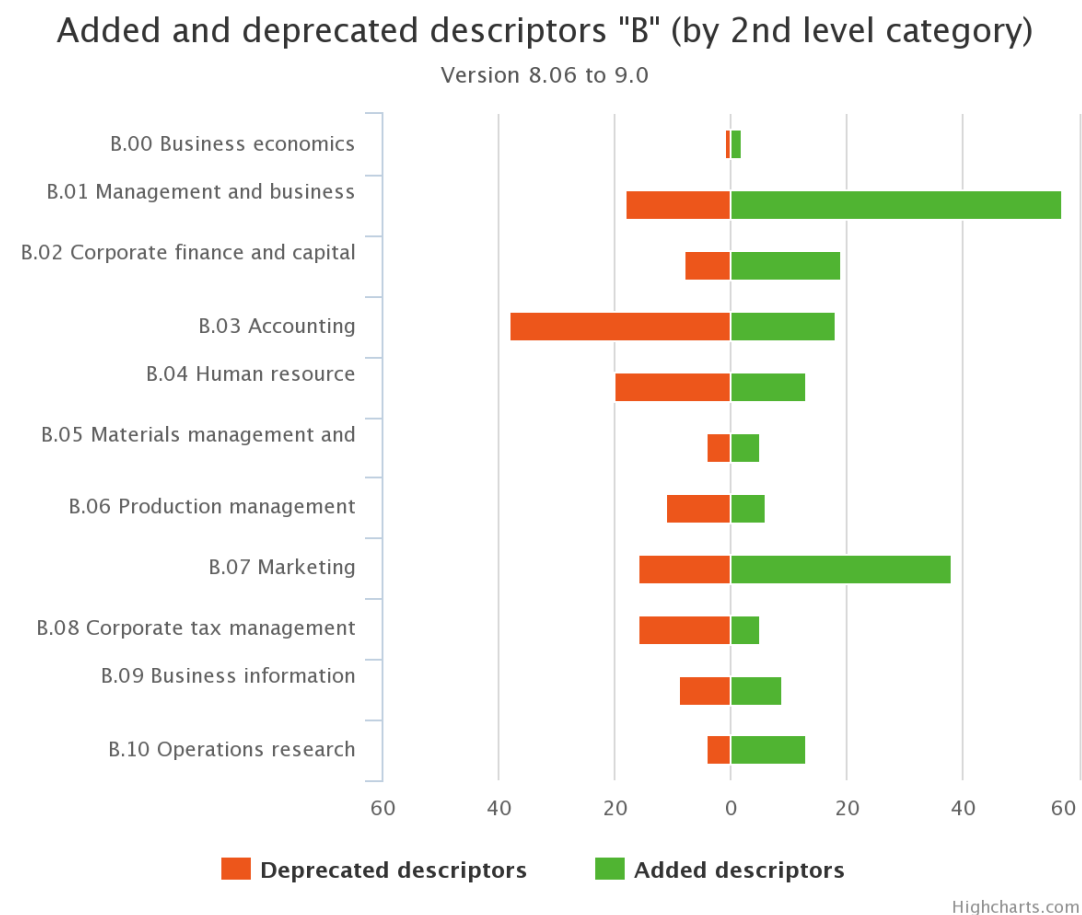


FIG. 5 Sub-thesaurus Business economics: Added and deprecated descriptors by 2nd level category, with drilldowns

6. Exposing the complete history of single concepts

From the point where a single concept is viewed, it would be very useful to be able to obtain the full history of that particular concept. To this end, a SPARQL construct query on the data of the version store has been developed²⁰. It focusses on a single RDF subject and builds a temporary RDF graph, grouping all triple insertions and deletions for this subject by version delta. The query makes use of the Delta ontology as introduced by (Berners-Lee & Connolly, 2009). That allows us to track changes in (preferred or alternate) labels as well as changes in the relationships structure. Preliminary “Concept history (RDF/Turtle)” links to this graph are included on the web pages for all STW concepts since version 9.0.²¹ It is planned to transform the concept version graph to a formatted web page with human-readable labels for the concepts in one of the upcoming versions of STW.

7. Future work

Besides providing a human-readable concept history, the emphasis for future work lies in the field of extending the described methodology to other SKOS vocabularies, and to probe and test it in various use case scenarios.

The methodology described in this paper is intended to work with any published SKOS vocabulary, without the need for out-of-band knowledge sealed in its maintenance environment and processes. First results show that it could be applied and worked for the Thesaurus for the Social Sciences (TheSoz), which differs from STW in the use of SKOS-XL labels. Further experiments are under way with the Finnish General Ontology (YSO), which makes heavy use of SKOS collections, and with the Agrovoc thesaurus maintained by FAO and available in multiple languages, which differ largely in coverage. We can assume that adapting to the different specifics of individual thesauri will reveal commonalities as well as fields where additional restrictions or extensions will prove necessary, as it has been shown above for the descriptor and subject category sub-types of concepts within STW.

The reports described in section 4 are currently intended and optimized for mostly human consumption. While they are provided in machine-readable JSON format, further work is required to evaluate their use in (semi-) automatic processing scenarios as described in the first chapter of this paper. This will reveal ways to bundle the data which are better suited to both fully automated update tasks as well as roll out of changes which require human judgment and are poorly supported by maintenance tool chains at the moment. Publication in “raw” RDF may be useful for merging data. For example, the set of concepts which had been split up could be merged with the number of times these concepts had been used for indexing local documents, in order to estimate the impact of a version update. Or the labels which have been moved to a new concept can be searched automatically within the titles or abstracts of all documents indexed with the split-up concept, in order to generate a list of suggestions for a semi-automatic re-indexing workflow.

The skos-history project should be instrumental in gathering information about differences in thesaurus and classification structures and different usage scenarios, in order to develop a set of tools and best practices to trace change in knowledge organization systems.

²⁰ https://github.com/jneubert/skos-history/blob/master/sparql/concept_deltas.rq

²¹ e.g., <http://zbw.eu/stw/version/9.0/descriptor/16269-4/about>

Acknowledgements

With thanks to Sini Pessala (National Library of Finland) and Johan De Smedt (Tenforce, Belgium) for the cooperation and the discussions in the context of the skos-history project, and to Manuela Gastmeyer (ex-ZBW), who has built and maintained STW over more than two decades, including the overhaul traced in this paper.

References

- Berners-Lee, T., & Connolly, D. (2009, August 27). Delta: an ontology for the distribution of differences between RDF graphs. RDF Diff, Patch, Update, and Sync -- Design Issues. Retrieved September 2, 2013, from <http://www.w3.org/DesignIssues/Diff>
- De Smedt, J., Vrang, M. le, & Papantoniou, A. (2015). ESCO: Towards a Semantic Web for the European Labour Market. Accepted for the WWW Workshop Linked Data on the Web, Florence, Italy.
- Hillmann, D. I., Sutton, S. A., Phipps, J., & Laundry, R. (2006). A Metadata Registry from Vocabularies UP: The NSDL Registry Project. *arXiv:cs/0605111*. Retrieved from <http://arxiv.org/abs/cs/0605111>
- ISO TC46/SC9/WG8, & Isaac, A. (2012, June). Correspondence between ISO 25964 and SKOS/SKOS-XL Models. Retrieved from http://www.niso.org/apps/group_public/download.php/9627/Correspondence%20ISO25964-SKOSXL-MADS-2012-10-21.pdf
- Miles, A., & Bechhofer, S. (2009, August 18). SKOS Simple Knowledge Organization System Reference. Retrieved March 23, 2015, from <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>
- Neubert, J. (2009). Bringing the “Thesaurus for Economics” on to the Web of Linked Data. In *Proc. WWW Workshop on Linked Data on the Web (LDOW 2009), Madrid, Spain*. Retrieved from http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf
- Neubert, J. (2014, November 14). Publishing SPARQL queries live. Retrieved from <http://zbw.eu/labs/en/blog/publishing-sparql-queries-live>
- Pessala, S., Seppälä, K., Suominen, O., Frosterus, M., Tuominen, J., & Hyvönen, E. (2011). MUTU: An Analysis Tool for Maintaining a System of Hierarchically Linked Ontologies. In *ISWC 2011 - Ontologies come of Age Workshop (OCAS)*. Bonn, Germany. Retrieved from <http://www.seco.tkk.fi/publications/2011/pessala-et-al-mutu-2011.pdf>
- Rietveld, L., & Hoekstra, R. (2015, under review). The YASGUI Family of SPARQL Clients. Retrieved from <http://www.semantic-web-journal.net/content/yasgui-family-sparql-clients>

The Linkable Neil Armstrong: Using BIBFRAME to Increase Visibility of Digital Collections

Carolyn Hansen
University of Cincinnati
United States
carolyn.hansen@uc.edu

Sean Crowe
University of Cincinnati
United States
sean.crowe@uc.edu

Abstract

This report describes the initial phase of an experimental project to increase Web visibility of the Neil Armstrong Commemorative Archive, a digital collection of archival materials concerning astronaut Neil Armstrong's tenure at the University of Cincinnati. The project description includes explanation of the mapping process from Qualified Dublin Core to BIBFRAME as well as data reconciliation and linking to external authorities such as id.loc.gov, VIAF, and Wikipedia. Next steps in the project, such as integrating related MARC datasets from local library catalogs, are also discussed.

Keywords: linked data, BIBFRAME, Dublin Core, metadata, digital collections

1. Introduction

Neil Armstrong, celebrated astronaut and the first person to walk on the moon, was also a professor of aerospace engineering at the University of Cincinnati (UC). In October 2013, the UC Libraries' Digital Collections and Repositories Department published the Neil Armstrong Commemorative Archive, a digital collection of unique archival materials concerning Armstrong's tenure at UC.¹ The collection contained two hundred and eighteen items, including letters, photographs, artifacts, and ephemera. Although the collection was extensively described using established information standards such as the Qualified Dublin Core (DC) metadata standard and Library of Congress Name and Subject Authority Headings (LCNAF and LCSH, respectively), its discoverability outside of library catalogs and repositories was limited by the structured metadata schemas that those systems required. In order to capitalize on the power of linked open data to improve the collection's visibility on the Web, an experimental project was undertaken by UC library faculty to map the original DC metadata to the Bibliographic Framework (BIBFRAME) data model, reconcile and link the data to external authorities using the OpenRefine application, and publish the data as expressed in the Resource Description Framework (RDF).

This report will describe the initial phase of the project, including explanation of the mapping process from Qualified Dublin Core to BIBFRAME as well as data reconciliation and linking to external authorities such as id.loc.gov, the Virtual International Authority File (VIAF), and Wikipedia. In addition, next steps in the project, such as integrating related MARC datasets from local library catalogs, will be discussed.

2. Methodology: Metadata for Discovery

Although data is often considered the unbiased product of research, the environment in which it is created and stored impacts its content, structure, and meaning. In this project, the original dataset consisted of Qualified DC records created for UC's DSpace repository, the Digital Resource Commons (DRC).² For purposes of this project, the original records were

¹ <https://drc.libraries.uc.edu/handle/2374.UC/713357>

² <https://drc.libraries.uc.edu/>

conceptualized as both metadata (abstract representations of digital objects) and data (a set of elements and values generated during the cataloging process). Viewing the metadata in the context of the larger dataset impacted the mapping approach from DC to BIBFRAME; specifically, a lossless migration between standards was not sought. Instead of focusing on comprehensive or archival mapping that preserved the authenticity and content of the original data in a one-to-one mapping, a flexible approach was taken in which a core set of properties (see Table 1) needed for discovery were identified and mapped.

TABLE 1. Discovery Metadata Mapping

Qualified DC (UC Armstrong Collection)	Simple DC ³	BIBFRAME Core Class	BIBFRAME Property
dc.contributor dc.contributor.author dc.contributor.photographer dc.contributor.other	dc.contributor	bf:Work bf:Authority	bf:contributor
dc.date.available	dc.date	bf:Instance	bf:providerDate
dc.identifier.uri	dc.identifier	bf:Instance or bf:Annotation ⁴	bf:uri
dc.publisher.digital	dc.publisher	bf:Instance	bf:providerName
dc.subject dc.subject.lcsh	dc.subject	bf:Work bf:Authority	bf:subject
dc.title	dc.title	bf:Work bf:Authority	bf:title
N/A	N/A	bf:Instance	bf:providerPlace

There are many benefits to a “metadata for discovery” approach.⁵ First, being able to omit properties from the mapping provides a clean dataset without idiosyncratic data. For example, UC’s DRC repository contained legacy data that conformed to outdated OhioLINK consortial practices (see examples of non-mapped properties from the original dataset in Appendix I); this data did not increase discoverability or add value in a linked data environment. Second, since BIBFRAME is an emerging model that is relatively unstable, eliminating properties that are not crucial for discovery reduces the amount of data cleanup needed as the model changes. Third, creating a lightweight dataset for discovery is time-efficient, allowing for mapping alterations to be made on the fly. Lastly, mapping for discovery simplifies working with multiple instances of physical objects and digital surrogates. Instead of accounting for the various instances of one work, focus can be placed on the digital surrogate. For example, the Armstrong dataset contained a digital surrogate for a photograph that had three instances in its lifecycle: it was created by the

³ For description of the 15 properties of Simple DC, see: <http://dublincore.org/documents/dces/>

⁴ The BIBFRAME model and vocabulary are still being defined and there is room for interpretation in how to conceptualize and map certain properties. In the Armstrong sample data, bf:uri property is entered under the bf:Instance class, but a case could also be made to enter it under bf:Annotation.

⁵ The authors acknowledge that this is only possible if there is an existing system to store and make the comprehensive records accessible.

photographer,⁶ published as a reproduction in a magazine, and published as a digital surrogate in the Neil Armstrong Commemorative Archive. Archival materials and museum objects often have multiple stages in their lifecycles, which can be difficult or cumbersome to express in BIBFRAME, since the objects differ from traditional forms of publication. By relying on existing platforms and specialized descriptive standards such as Encoded Archival Description (EAD) or the Visual Resources Association (VRA) standard for comprehensive description, BIBFRAME mappings can be simplified. This is a significant distinction; to paraphrase Nancy Fallgren, Metadata Specialist Librarian at the National Library of Medicine (NLM), “MARC became a descriptive scheme in addition to an encoding standard. We should not do that with BIBFRAME.”⁷

3. Mapping Dublin Core to BIBFRAME

BIBFRAME was initially created as a replacement standard for MARC, but it has been advertised as a more inclusive model that can accommodate a broader user community. This may be true in the future; however, working with BIBFRAME outside of text-based materials and MARC record migration is challenging in the current environment. Part of the problem is that this work is very new and there are few example datasets available from BIBFRAME early adopters. The datasets that are available via LC’s BIBFRAME website⁸ all focus on MARC record migration using LC’s Transformation Tool. For those working with non-MARC metadata, digital collections, or archival materials described at the item level, these datasets are of limited assistance. To the authors’ knowledge, the mapping process described in this report is the first to work with BIBFRAME and non-MARC metadata in digital collections.

The first step in the mapping process was to eliminate idiosyncratic properties from the dataset. For example, the DC properties referring to events in the lifecycle of the physical object such as `dc.date.created` were removed (see Appendix I for a full list of unmapped properties). Then, the remaining properties that could not be expressed in BIBFRAME as Uniform Resource Identifiers (URIs) were isolated and examined. If these properties provided information that was structurally important for the BIBFRAME core classes, they were retained.⁹ Next, the remaining DC properties were mapped to the four core BIBFRAME classes: Work, Instance, Annotation, and Authority (see Table 1.1 for discovery metadata mapping and Table 1.2 for visual representation of the BIBFRAME model). Initial mapping to the core classes helped to intellectually organize the data; this was important when working with BIBFRAME data serialized as RDF because the heavy use of URIs made the RDF difficult to read. Finally, the DC properties were mapped to corresponding BIBFRAME properties. As a result of the “metadata for discovery” approach in the mapping and the work-centric nature of BIBFRAME, this project used few properties that mapped to BIBFRAME Instance or BIBFRAME Annotation. In this dataset, BIBFRAME Instance was used to describe the publication of the digital surrogate on UC’s DRC Repository; it was not used to represent earlier publications or other events in the object’s lifecycle (for more information on describing archival materials and digital surrogates, see Section 2).

⁶ The act of creation could be considered two separate instances. The physical act of taking the photograph could be one instance, and the development of the film into a print would be the second. For simplicity, the authors define this as one instance.

⁷ This is paraphrased from Nancy Fallgren’s presentation “Experiments in BIBFRAME: A Modular Approach.” given at the American Library Association Midwinter Meeting in January 2015.

⁸ <http://www.loc.gov/bibframe/implementation/>

⁹ Currently, www.bibframe.org does not specify input requirements for properties.

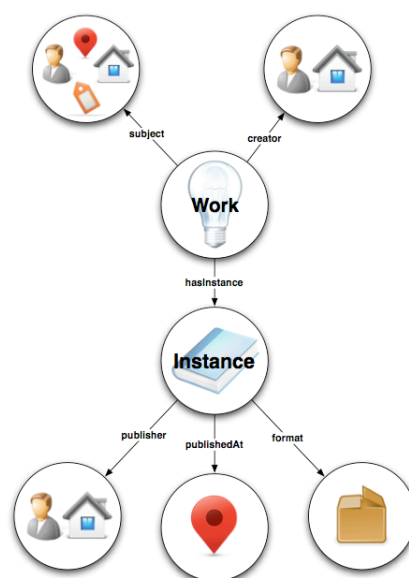


FIG. 1. "BIBFRAME model" by Zepheira, under contract from the U.S. Library of Congress - <http://www.loc.gov/bibframe/docs/images/bibframe.png>. Licensed under Public Domain via Wikimedia Commons http://commons.wikimedia.org/wiki/File:BIBFRAME_model.png#/media/File:BIBFRAME_model.png

4. Defining Authorities

In traditional description, authorities refer to controlled vocabularies that bring together variant forms of a name for people, organizations, subjects, etc. The BIBFRAME concept of authority is more flexible; it is defined only as “Representation of a key concept or thing.”¹⁰ In practice, this representation is expressed as both strings (bf:label, bf:titleValue, bf:authorizedAccessPoint) and things (bf:creator, bf:subject, bf:title) in the form of URIs. There are currently no guidelines or best practices regarding what constitutes a reliable authority in terms of site content, although stable URIs are needed from the technical perspective. In this project, LC authorities were contained in the original DC dataset, so id.loc.gov was used as the primary authority (bf:hasAuthority). Reference authorities (bf:referenceAuthority) included VIAF, Wikipedia, and organizational websites for corporate bodies (see Table 2 for authority mapping).

TABLE 2. Authority Mapping

BIBFRAME Property	Authority Used
bf:hasAuthority	LCNAF; LCSH
bf:referenceAuthority	VIAF; Wikipedia; organizational websites

5. Linking and Data Reconciliation

The authors investigated several enrichment services for reconciling the core metadata set. Since the DC metadata included LC subject headings and names, the team first explored the possibility of using id.loc.gov for reconciliation; unfortunately, there was no SPARQL endpoint or other batch query interface for id.loc.gov that could be used. As an attempted workaround, the

¹⁰ <http://bibframe.org/vocab-list/#Authority>

authors downloaded the LC name authority file and using Apache Jena tools,¹¹ loaded the file into a TDB, to spin up a local SPARQL endpoint for name reconciliation. The LC name authority file was sizable (> 30 GB); even with local access to query the TDB, name reconciliation for one column with ~70 unique entries had a runtime in excess of three hours.

Since the results were not optimal and in the interest of time, the authors manually created BIBFRAME authority objects and included links from several enrichment services such as VIAF, id.loc.gov, and Wikipedia. The BIBFRAME authorities were then integrated with the core dataset using the OpenRefine reconciliation function to link the separate files. In an ideal process, there would be a reconciliation service for id.loc.gov, since much of the legacy metadata for the DRC dataset included vocabularies and authorities from LC. However, even if a SPARQL endpoint was available, id.loc.gov does not contain a complete dataset of LCNAF and LCSH. Problems also arise when a URI is available for the parent body of an organization, but not the subordinate body as found in the local dataset. For this project, the authors linked to parent bodies, even though the matches were not exact (see Table 3 for examples). This approach also worked for LCSH subject strings when a primary topical heading had an authority but the string did not.

TABLE 3. Example of id.loc.gov partial matches to DC dataset

Entity From DC Dataset	Partial Match (id.loc.gov)
American Institute of Aeronautics and Astronautics. Student Chapter	http://id.loc.gov/authorities/names/n79053067 (parent body)
University of Cincinnati. Board of Trustees	http://id.loc.gov/authorities/names/n79034519 (parent body)

6. Transformation Process

One of the goals of this project was to develop scripts for conversion of data from DC to BIBFRAME. Encoding the conversion process into scripts offered the advantage of reuse and easy adaptation. However, faced with challenges in conceptualizing the process and in the interest of experimentation, much of the work was done manually for this first phase. The authors chose to focus on outlining the model and closely curating a small dataset (218 records) as a proof of concept. The input data consisted of a blend of DC metadata in CSV format and manually created RDF/XML. The DC metadata comprised the foundation of the dataset, augmented and linked with additional, hand-curated BIBFRAME elements in RDF/XML.

The output dataset was self-contained, comprised of BIBFRAME Work, Instance, and Authority data, concatenated from separate files. The BIBFRAME Works were mapped directly from the DC dataset; each type was generated with OpenRefine and the Digital Enterprise Research Institute (DERI)'s RDF extension. The DERI RDF extension¹² includes RDF skeleton functionality to map data to namespaced elements for export. For this project, a custom skeleton based on the discovery metadata map was created.

7. Next Steps and Recommendations

The vision for this project is to package, publish, and optimize linked data for all collections at UC Libraries. The authors agree with the philosophy of the LibHub initiative¹³ in describing

¹¹ <https://jena.apache.org/>

¹² <http://refine.deri.ie/>

¹³ <http://www.libhub.org/>

efforts to use the Semantic Web and Search Engine Optimization (SEO) techniques to better position cultural heritage institutions for discovery via commercial Web search engines. Achievement of this vision will require several subsequent steps.

1. Investigate other enrichment services for streamlining of the reconciliation process. For example, using OCLC FAST for subject authorities and ISNI for name authorities.
2. Map other ontologies in the dataset; ex. FOAF, SKOS, etc.
3. Server space - For further experimentation, the department has procured a public-facing virtual server for hosting local linked data sets on an ongoing basis and has plans to post linked data sets for public consumption.
4. Process MARC records for UC Libraries' physical collections into BIBFRAME and link with special collections datasets to improve discovery. Ultimately, we will want to take steps to review systems, enterprise-wide, and assess fitness for modeling and exposing linked data. Where possible, integrate linked data publishing at the system level and implement tools for working with linked data natively.

8. Conclusion

This project represents the first linked data initiative for UC Libraries. The authors spent time experimenting with tools and technologies to convert and reconcile legacy metadata for a high-interest special collection. Emerging trends in library linked data and the Semantic Web are central to several of the UC Libraries' strategic initiatives; touching on issues of access, discovery and preservation. Libraries house a wealth of data in many formats, most of which, because of structure or format, are not easily adapted for linking and sharing on the Web. The BIBFRAME initiative offers a core standard for expressing MARC but remains flexible enough to encompass other flavors of metadata. The task of migrating Qualified Dublin Core to BIBFRAME, even with the loosened constraints of our focus on discovery rather than comprehensive representation, is a demonstration of that flexibility. Although letting go of traditional ideas about metadata and description is difficult, thinking in terms of system needs for successful identification and linking of data is an essential step to discovery.

Acknowledgements

The authors would like to thank the following individuals for their help and support for this project: Elna Saxton, Head of Content Services, University of Cincinnati; Leslie Schick, Associate Dean of Library Services & Director of the Health Sciences Library, University of Cincinnati; and Victoria Mueller, Senior Information Architect and Systems Librarian, Zepheira, LLC.

Bibliography

- Dublin Core Metadata Initiative (2012). *Dublin Core Metadata Element Set, Version 1.1*. Retrieved from <http://dublincore.org/documents/dces/>.
- Fallgren, Nancy (2015). *Experiments in BIBFRAME: A Modular Approach*. Retrieved from <http://connect.ala.org/node/68263>.
- Library of Congress (2015). *Bibliographic Framework Initiative*. Retrieved from <http://bibframe.org>.
- University of Cincinnati Libraries (2013). *Neil Armstrong Commemorative Archive*. Retrieved from <https://drc.libraries.uc.edu/handle/2374.UC/713357>.

Appendix I: Properties From Original Dataset Not Mapped to BIBFRAME

Qualified DC (UC Armstrong Collection)	SIMPLE DC
dc.date.created <i>Note:</i> Refers to original object, not digital surrogate	dc.date
dc.date.digitized <i>Note:</i> date.available was used as publication date for digital surrogate	dc.date
dc.description	dc.description
dc.format	dc.format
dc.language.iso	dc.language
dc.publisher	dc.publisher
dc.publisher.OLinstitution <i>Note:</i> Legacy OhioLINK property	dc.publisher
dc.relationispartof	dc.relation
dc.relationispartofseries	dc.relation
dc.relation.uri	dc.relation
dc.rights	dc.rights
dc.rights.uri	dc.rights
dc.source	dc.source

Appendix II: Sample Data Serialized as Turtle

```

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bf: <http://bibframe.org/vocab/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://data.libraries.uc.edu/armstrong/works/211>
  a bf:Work , bf:Text ;
  bf:contributor <http://data.libraries.uc.edu/armstrong/bibframe/people/52> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/46> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/3> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/26> ;
  bf:subject <http://data.libraries.uc.edu/armstrong/bibframe/people/52> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/3> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/26> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/46> ;
  bf:title <http://data.libraries.uc.edu/armstrong/bibframe/titles/212> ;
  bf:uri <http://hdl.handle.net/2374.UC/731329> .

<http://data.libraries.uc.edu/armstrong/bibframe/instances/211>
  a "http://bibframe.org/vocab/Electronic" ;
  a bf:Instance ;
  bf:instanceOf "Working proposal related to individualized oxygen systems and artificial

```

```

organs, April 14, 1977" ;
  bf:instanceTitle "Working proposal related to individualized oxygen systems and artificial
organs, April 14, 1977" ;
  bf:provider "University of Cincinnati. University of Cincinnati Libraries" ;
  bf:providerDate "2013" ;
  bf:providerPlace "Cincinnati, Ohio" ;
  bf:uri "http://hdl.handle.net/2374.UC/731329" .

<http://data.libraries.uc.edu/armstrong/bibframe/titles/212>
  a bf:Title ;
  bf:AuthorizedAccessPoint "Working proposal related to individualized oxygen systems and
artificial organs, April 14, 1977" ;
  bf:titleValue "Working proposal related to individualized oxygen systems and
artificial organs, April 14, 1977" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/3>
  a bf:Person ;
  bf:AuthorizedAccessPoint "Armstrong, Neil, 1930-2012" ;
  bf:hasAuthority "http://id.loc.gov/authorities/names/n80008815" ;
  bf:label "Armstrong, Neil, 1930-2012" ;
  bf:referenceAuthority "http://viaf.org/viaf/111826406" ,
"http://en.wikipedia.org/w/index.php?title=Neil_Armstrong&oldid=650449902" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/26>
  a bf:Person ;
  bf:AuthorizedAccessPoint "Heimlich, Henry J." ;
  bf:hasAuthority "http://id.loc.gov/authorities/names/n79107850" ;
  bf:label "Heimlich, Henry J." ;
  bf:referenceAuthority "http://viaf.org/viaf/269976816" ,
"http://en.wikipedia.org/w/index.php?title=Henry_Heimlich&oldid=643760119" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/52>
  a bf:Person ;
  bf:AuthorizedAccessPoint "Rieveschl, George, 1916-2007" ;
  bf:hasAuthority "http://id.loc.gov/authorities/names/no98002197" ;
  bf:label "Rieveschl, George, 1916-2007" ;
  bf:referenceAuthority "http://viaf.org/viaf/26675176" ,
"http://en.wikipedia.org/w/index.php?title=George_Rieveschl&oldid=577487552" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/46>
  a bf:Person ;
  bf:AuthorizedAccessPoint "Patrick, Edward A., 1937-" ;
  bf:hasAuthority "http://id.loc.gov/authorities/names/n85114241" ;
  bf:label "Patrick, Edward A., 1937-" ;
  bf:referenceAuthority "http://viaf.org/viaf/109256464" .

```




Peer-Reviewed Posters

EZID: Easy Identifier and Metadata Management

John Kunze
University of California,
California Digital Library
USA
jak@ucop.edu

Greg Janée
University of California,
California Digital Library
USA
gjanee@ucop.edu

Joan Starr
University of California,
California Digital Library
USA
joan.starr@ucop.edu

Keywords: persistent identifiers; metadata; resolvers; API; sustainability; N2T; DOI; ARK

Abstract

EZID (pronounced easy-eye-dee at ezid.cdlib.org) is an innovative service supporting the creation and management of identifiers, their accompanying metadata, and long-term access to things on the Internet. It is one of the few services that can supply a diversity of identifier and metadata types, and do so at the earliest stages of content development, long before the content is archived or its value is understood.

EZID is run by a team within the California Digital Library (CDL), which serves the libraries of the ten campuses of the University of California, partners with national libraries, maintains the ARK identifier scheme, and belongs to global identifier organizations such as DataCite and CrossRef (FIG. 1). Started in 2010, EZID now has over 100 customers on three continents and users on all continents. In fact it is the largest and fastest growing member of the DataCite consortium. The EZID user interface is currently being revised to support multiple languages.

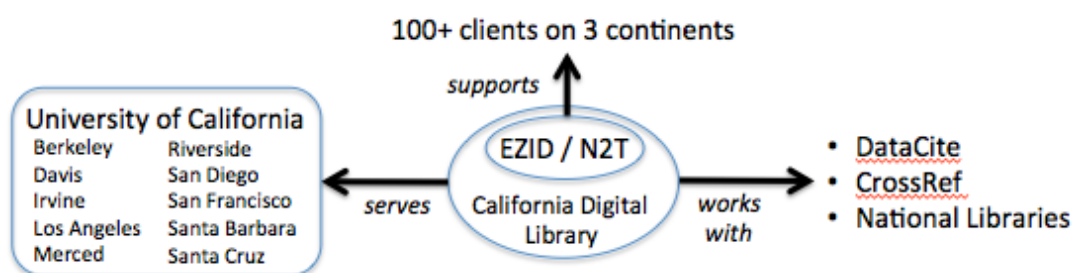


FIG. 1. Organizational context of EZID at the CDL.

EZID is unusual in supporting different kinds of identifiers. Its identifiers and metadata can describe anything of any type: documents, films, digitized maps, datasets, fossils, stars, vocabulary terms, people, etc., and it supports any identifier scheme (currently ARKs and DOIs), as well as a variety of metadata profiles, such as Dublin Core, Kernel, and DataCite.

Also unusual is that identifiers may be used for objects that are still under development. An EZID client can create “reserved” identifiers that are held privately until, for example, a draft manuscript citing them will be published. A demo mode allows anyone (no login required) to create fully functional temporary identifiers. Clients can create and use “preservation-ready” identifiers for objects that are incubating or speculative; such objects need not receive a new name when (or indeed if) they are officially published or archived, perhaps years later.

While any URL can be made persistent by carefully managing a local web server and its redirection tables, some organizations need help doing this. EZID provides them with both a user interface and an API (application programming interface) to make centralized metadata

management easy, secure, and automatable. Every identifier has an authorized maintainer (transferable, for example, to a successor organization) and a profile (a metadata element) guiding how all of its metadata will be presented for display, crosswalking, indexing, etc. EZID manages DOIs and ARKs that are tracked in Thomson Reuters' Data Citation IndexSM.

Persistent identifiers that work with web browsers are actually URLs with carefully chosen hostnames. Sometimes a hostname identifies a "resolver", which is a special web server that forwards (redirects) public internet access requests to an object's current location, as recorded in the identifier's metadata. EZID uses two resolvers – the hostnames in these identifier examples:

http://doi.org/10.5072/FK234567	<i>a DOI identifier</i>
http://n2t.net/ark:/99999/fk456789	<i>an ARK identifier</i>

These affiliated resolvers, doi.org and n2t.net, support persistent identifier reference for any Internet user. EZID is one of the services, along with data centers and publishers, that updates DOIs at doi.org. Along with the Internet Archive, EZID also updates ARKs at n2t.net.

The N2T (Name-to-Thing) resolver at n2t.net net is non-traditional. The traditional approach to identifier persistence has been to develop a new identifier scheme and lock it up with redirection and management services designed exclusively for it. Thus the PURL, Handle, DOI, and URN schemes each has its own service "silo", and much duplicative software to manage, redirect, check links, etc. In contrast, N2T serves identifiers of any type (currently ARKs and DOIs). It is open, scalable infrastructure implemented from scratch using simple open source packages.

Traditional scheme-specific silos raise concerns for open access. With DOIs for example, it happens that any one of three specific service organizations could in theory insert advertising in or even shut down access to all mainstream scholarly journal content. EZID and N2T are deliberately scheme-agnostic, and N2T was envisioned as a resolver that could be maintained in perpetuity by a consortium of memory organizations. N2T is scalable infrastructure currently homed at the CDL and high availability is one reason CDL recently began running its infrastructure in the Amazon cloud. Until global replication across multiple regions is achieved, CDL continues to partner with EDINA to maintain an N2T replica in Edinburgh, UK.

N2T has a unique feature called "suffix passthrough" that permits one identifier for a complex object to enable resolution for many thousands of component sub-identifiers, which greatly reduces the identifier maintenance burden. Planned features in support of open linked data (semantic web) applications include "content negotiation" and a powerful inflection mechanism (short standardized extensions added to the end of an identifier).

With a view to sustainability, EZID charges a small annual fee to recover costs. Persistence is a priority, so clients that can no longer pay the fee nonetheless still retain login privileges in order to continue managing their existing identifiers. Customers include libraries, museums, archives, government agencies, publishers, and commercial data services. N2T resolver sustainability is a separate but equally important concern.

Using Metadata for Interoperability of Species Distribution Models

Cleverton Ferreira Borba
University of Sao Paulo, Brazil
UNASP, Brazil
cleverton.borba@gmail.com

Pedro Luiz Pizzigatti Corrêa
University of Sao Paulo,
Brazil
pedro.correa@usp.br

Keywords: Metadata, Species Distribution Modeling, Dublin Core, Darwin Core, architecture, interoperability.

1. Use of Metadata for Species Distribution Modeling (SDM)

This poster presents the use of metadata patterns for the field of Species Distribution Modeling and also presents a proposal for application of metadata to ensure interoperability between models generated by tools of Species Distribution Modeling (SDM).

According to Peterson et al. (2010) the area of Biodiversity Informatics is responsible, by the use of new technologies and computational tools, to meet the demand for support the biodiversity conservation. Portals of biodiversity, taxonomic databases, SDM tools help the scientists and researchers to decide the best for the biodiversity conservation. However, Berendsohn et al. (2011) says that one of the most serious problems in scientific biodiversity science is the need to integrate data from different sources, software applications and services for analysis, visualization and publication and thus offer an interoperability of data, information, application and tools.

In this context, the metadata patterns available, has been used to help the scientists and researchers to define vocabulary and data structure for analysis, visualization and publication of biodiversity data. Examples of metadata used in SDM are: Dublin Core (DC) (DCMI, 2012), Darwin Core (DwC) (Wieczorek *et al.*, 2012), Darwin Core Archive (DwC-A) (GBIF, 2010), EML – Ecological Metadata Language (Fegraus *et al.*, 2005), etc.

Biodiversity portals like GBIF (Global Biodiversity Information Facility - <http://www.gbif.org/>), ALA (Atlas of Living Australia – <http://www.ala.org.au/>), Specieslink (<http://splink.cria.org.br/>) also use metadata standards to support the data integrity, interoperability, and the data standardization. SDM tools, that use the data provided by this portals, to produce species distribution models, also support the metadata domain for their proposal.

Based on this information an application of metadata to ensure interoperability between models of SDM is presented below.

2. Application of Metadata for Ensure Interoperability between Models of Species Distribution Modeling

To support and ensure the interoperability between models generated by SDM tools, we propose the use of the DC and DwC metadata. The metadata information generate should have the minimum data for reuse in the same SDM tool or another one. The XML archive contain the biodiversity data used (occurrence points [presence/absence]), the algorithm and the parameter used for calculate the model, climatic package, and the model map provided for the SDM tool. The Figure 1 shows an example that how each metadata pattern contributes with the SDM domain in this study.

Wieczorek *et al.* (2012) says that “an essential step towards understanding global patterns of biodiversity is to provide a standardized view of these heterogeneous data sources to improve interoperability”, and that is the object to apply the metadata for Models of SDM.

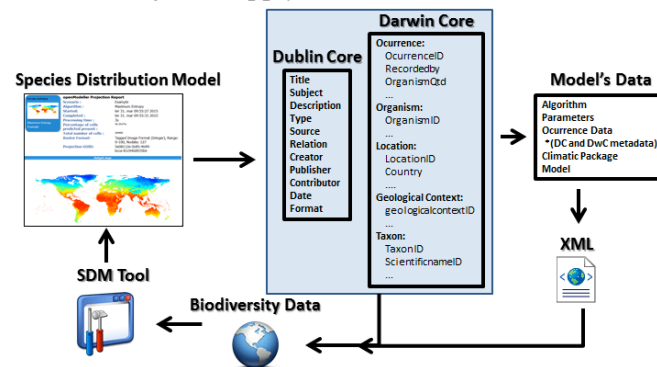


FIG. 1 - Use of Metadata for Species Distribution Modeling

The Figure 2 shows the proposed architecture for the SDM tools where the XML archive generated by a plug-in installed, has a metadata pattern to turn available the information used to make the model of species distribution (like algorithms, biodiversity data, parameters, climatic packages, etc.).

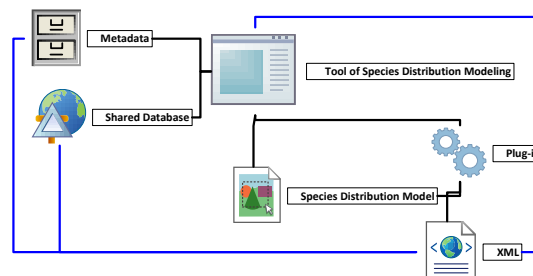


FIG. 2 - Proposed the use of metadata standards for species distribution modeling tools.

3. Conclusion

Through this research is possible to identify the importance of the metadata for the Biodiversity Informatics, specifically for Species Distribution Modeling. Just using a metadata is possible the interoperability between biodiversity data. In this study we proposed the use of metadata pattern for generate models of SDM tools through the development of a plug-in that presents a XML archive based in DC and DwC metadata to be reused, or offered in a portal of biodiversity.

As future work, we suggest the analysis with other metadata patterns and the use of the JSON archive for exportation of the model's data of species distribution.

References

- Berendsohn, W. G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K. & Müller, A. (2011) Biodiversity information platforms: From standards to interoperability. *ZooKeys*, v. 150, p. 71-87.
- DCMI. (2012). Dublin Core Metadata Element Set, version 1.1: Reference description. Retrieved March 31, 2015, from <http://dublincore.org/documents/dces/>
- Fegraus, E. H., Sandy, A., Matthew, J. B. & Mark, S. (2005) Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America*, p. 158-168.
- GBIF. (2010) Darwin Core Archive – How-to Guide. Copenhagen: Global Biodiversity Information Facility.

- Peterson A. T., Knapp S., Guralnick R., Soberón J. & Holder M. (2010) Perspective: The big question for biodiversity informatics. *Systematics and Biodiversity*. The Natural History Museum. 8(2) 159-168.
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7.

Interlinking Two Institutional KOS about Agroecology: Using LOD Agrovoc to Circumvent the Language Barrier in Identifying Terminological Intersections

Sophie Aubin
INRA, France
sophie.aubin@versailles.inra.fr

Pascal Aventurier
INRA, France
pascal.aventurier@avignon.inra.fr

Ivo Pierozzi Júnior
Embrapa, Brazil
ivo.pierozzi@embrapa.br

Leandro H. M. Oliveira
Embrapa, Brazil
leandro.oliveira@embrapa.br

Keywords: semantic interoperability; vocabulary alignment; open linked data; metadata; agroecology , skos.

1. Context and Aims of the work

INRA and Embrapa (respectively the French and the Brazilian national institutes for agricultural research) are historical partners in initiatives for knowledge and information management. Given the challenges involved in the mutual sharing of their technical-scientific production especially considering language barriers, efforts have been made to develop semantic interoperability between repositories and bibliographic databases of both institutions. INRA and Embrapa databases (respectively ProdINRA and BDP@) expose bibliographic data with Dublin Core, so the focus of this work was on dc:subject that aims at leveraging by a better semantic interoperability of vocabularies associated with these databases.

Among diverse agricultural subdomains, Agroecology is taking an increasingly important place in the issue of feeding the world, taking into account farmers activity, climate change, and agricultural modernization. Yet, each country and organization has a different understanding of Agroecology and what it covers exactly in terms of social issues, techniques, inputs, and for instance its relation to organic farming. So, considering the ubiquity as well as the ambiguity surrounding the subject, Agroecology was chosen as a case study since both institutions have strategic interests to develop and implement technological facilities to maintain specific terminologies while sharing mutual information. This scenario is extremely timely and demands a quick solution.

This work describes the methodological approach proposed to resolve the matter of indicating equivalent terms in both languages to the same concept recorded in Agrovoc related to the discipline of Agroecology. French and Brazilian vocabularies were not compiled using the same methods and then the analysis was not conducted similarly, requiring different treatment for each vocabulary until the Agrovoc SKOS exact match could be performed.

2. Material and Methods

INRA and Embrapa mutual collaboration aims to share information and knowledge according to their respective technical and methodological aptness. The Semantic Web, with its representation standards and tools, appears to be an interesting meeting point. More specifically, Agrovoc Linked Open Data (Agrovoc LOD) serialized in RDF SKOS and offering concept labels in Portuguese and French was chosen as the key solution. The building of the KOS (Knowledge Organization System) subsets was different for the two institutions.

INRA compiled the list of 3,140 French terms from VocINRA (INRA's own vocabulary) that were used to manually index 2,145 publications about Agroecology in the institutional repository ProdINRA. Onagui (Mazuel and Charlet, 2010), an open source tool designed to help alignment

of vocabularies in SKOS or OWL, was used to align the VocINRA concepts with those in Agrovoc.

Embrapa compiled a Brazilian Portuguese scientific textual corpus from 260 full papers about Agroecology, corresponding to 2,336,287 words and then performed a semi-automatic term extraction and a term matching using a specific tool developed to compare and reuse terminologies and conceptual structures from other KOS (Pierozzi Júnior et al., 2014). From the corpus a preliminary term list was built by semi-automatic term extraction and then it was matched with both Thesagro (a Brazilian Portuguese thesaurus) and Agrovoc-PT, producing a second term list where the exact match terms found in each of the two thesauri were identified and separated and translated in SKOS. The SKOS information from Agrovoc was further retrieved for those terms found at the same time in Thesagro and in Agrovoc.

3. Results

Out of the 3,140 selected French (FR) terms, 1,542 were found in Agrovoc LOD, on the bases of Stoilos (Stoilos et al. 2005) and Levenshtein distance algorithms implemented in Onagui with a nearly exact match distance (0.97). Results using these two string metrics to process the data are probably the same. There is no error in the alignment because it has a chosen value close to the exact match, that is 100% of the words in common. The chosen value for alignment was close to an exact match because the two thesauri are so big that we could not check the alignment values that were too far from the exact match.

Concerning the Brazilian Portuguese (PT/BR) Agroecology vocabulary, the preliminary term list was made up of 783,817 term candidates; the matching with both Agrovoc and Thesagro thesauri resulted respectively in 2,718 and 3,807 terms; exact SKOS match from Agrovoc resulted in 1,699 terms.

The intersection between the FR and PT/BR vocabularies totalizes 939 common URIs from Agrovoc LOD. Some keywords in common are: public health, rural development, recycling, technology transfer, *Raphanus sativus*, root nodulation, soil fertility, sowing depth, tropical climate. Some keywords are specific to INRA research domains as: vineyards, selective grazing, cauliflowers, agrosilvopastoral systems, *Dactylis glomerata*, environmental control. Finally, other keywords are specific to Embrapa: *Araucaria angustifolia*, *Jacaranda*, urban population, molasses, forest inventories, social indicators, passion fruits, root systems, *Leucaena leucocephala*, for example. INRA and Embrapa prepare the results from this work to be publicly available by its webservice.

4. Conclusions and perspectives

The primary motivation for collating INRA and Embrapa methodologies of building and using vocabularies was to implement faster and better semantic interoperability of the KOS used to index and thus share the large amount of scientific knowledge produced in France and Brazil. Agrovoc LOD was chosen to identify and map common terms (and consequently concepts) in the context of French and Brazilian agricultural knowledge using Agroecology as a study case.

Agrovoc LOD proved to be an interesting and feasible solution functioning as a pivot where the methodological differences in the construction of both vocabularies do not interfere in the final result, allowing both (1) the identification of already common terms used for the two institutions and (2) a set of specific terms in each language that might be incorporated into each other's vocabularies. The alignment between the INRA and Embrapa vocabularies prepares these two vocabularies to be linked when published in LOD. Documents contained in the respective institutional repositories of both organizations as well as documents from other institutions around the world, may then be found from the linked data to the Agrovoc URI of a specific term. This work also highlights some difficulties in the translation of certain terms in Agrovoc which will improve this multilingual vocabulary.

This conciliatory methodological model can be strengthened and systematized so that Embrapa and INRA can consider their contribution to broader initiatives in the agricultural domain like the project for a Global Agricultural Concept Scheme (Baker and Suominen, 2014).

References

Agrovoc LOD: <http://aims.fao.org/standards/agrovoc/linked-open-data>

Baker, Thomas and O. Suominen. Global Agricultural Concept Scheme (GACS): A multilingual thesaurus hub for Linked Data. 2014 http://aims.fao.org/sites/default/files/posts/attachments/GACS_Integration_Proposal_1.0_3.pdf

Mazuel Laurent and Jean Charlet "SPIM-AlignmentGUI - un logiciel d'aide à la réalisation d'alignements entre ontologies 2009. Inria http://ic2009.inria.fr/docs/posters/MazuelCharlet_Poster_IC2009.pdf

Pierozzi, Ivo Júnior, Marcia Izabel Fugisawa Souza, Tércia Zavaglia Torres, Leandro Henrique Mendonça de Oliveira and Leonardo Ribeiro Queiros. Gestão da informação e do conhecimento. In: Tecnologias da informação e comunicação e suas relações com a agricultura. Brasília, DF: Embrapa, 2014. Cap. 12. p. 237-260. URL: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/119627/1/capitulo12-085-14.pdf>

OnAGUI - Ontology Alignment GUI :<http://sourceforge.net/projects/onagui/>

Stoilos Giorgos;Stamou, Giorgos; Kollias, Stefanos (2005). A String Metric for Ontology Alignment. The Semantic Web – ISWC 2005 . Lecture Notes in Computer Science Volume 3729, pp 624-637

Dublin Core and CIDOC CRM Harmonization

Laís Carrasco
Unesp, Brazil
laiscarrasco@hotmail.com

Silvana A. Borsetti Gregorio Vidotti
Unesp, Brazil
vidotti@marilia.unesp.br

Keywords: Dublin Core Metadata Element Set (DCMES); CIDOC CRM; ontologies, metadata, information integration, semantics mappings.

1. Introduction

In order to integrate information from heterogeneous sources, ontologies as semantic technologies are a recommend solution. “An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents”. (Gruber, 2001) CIDOC Conceptual Reference Model (CIDOC CRM) is a very prominent ontology used for such purposes.

The CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. [...] In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives. (CIDOC CRM)

As semantics mapping can be a solution for information integration and Dublin Core is the most prominent metadata used to describe web resources, we propose a harmonization between Dublin Core and CIDOC CRM ontology. According to Nilsson (2010, p. 107) harmonized standards is “*a set of metadata standards that can be semantically embedded into another standard*”. Here, CIDOC CRM is used as the mediated schema to integrate metadata sources in the Cultural Heritage domain. It's important to mention that other works are making efforts in this direction, for example, the *Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM* headed by Doerr (2000).

2. Mapping Dublin Core into CIDOC CRM ontology

Beneath we present a semantic mapping from the Dublin Core Metadata Element Set (DCMES) into CIDOC Conceptual Reference Model entities in other to provide information integration.

TABLE 1: DCMES and CIDOC CRM Harmonization.

Dublin Core	CIDOC CRM	Dublin Core	CIDOC CRM
Contributor	E39 Actor E74 Group E41 Appellation E10 Transfer of Custody E66 Formation	Type	E55 Type E17 Type Assignment
Coverage	E50 Date E52 Time-Span E53 Place E47 Spatial Coordinates E45 Address E48 Place Name	Publisher	E12 Production E29 Design or Procedure E51 Contact Point
Creator	E39 Actor E40 Legal Body	Identifier	E42 Object Identifier E15 Identifier Assignment

	E66 Formation E74 Group E41 Appellation		E73 Information Object E71 Man-Made Stuff E70 Stuff
Language	E56 Language	Type	E55 Type E17 Type Assignment
Description	E5 Event E7 Activity E12 Production E14 Condition Assessment E3 Condition State E18 Physical Stuff E19 Physical Object E20 Biological Object E22 Man-Made Object E23 Iconographic Object E24 Physical Man-Made Stuff E25 Man-Made Feature E26 Physical Feature E28 Conceptual Object	Date	E2 Temporal Entity E4 Period E50 Date
		Rights	E40 Legal Body E30 Right E72 Legal Object
		Source	E42 Object Identifier E62 String E73 Information Object
		Format	E16 Measurement E29 Design or Procedure E54 Dimension E57 Material E58 Measurement Unit
		Subject	E73 Information Object E46 Section Definition
		Relation	E27 Site E31 Document

3. Final considerations

According to the literature, there are many XML metadata mapping to the CIDOC CRM ontology efforts, since this ontology is considered one of the most appropriate models in integration architecture. On the other hand, Dublin Core is the most used metadata in semantic web applications. In this way, metadata can be mapped into an ontology to provide interoperability of its data and to achieve information integration. When the different kind of metadata are mapped into an ontology the system can interoperate and the information access is higher as well as their information retrieval.

The major difficulty found in this research was that the Dublin Core Element Set has just 15 attributes, on the other hand, CIDOC CRM has 93 entities, making it difficult to express all CRM relationships, so in this work, we chose only those entities that have their concepts more similar to the DCMES.

As DCMES is the most prominent metadata used to describe web resources, a DCMES and CIDOC CRM cross-walking model will be developed in a future work in order to handle cultural heritage data representation into the web.

References

- CIDOC CRM. <http://www.cidoc-crm.org/>.
- Crofts et al. (2015). Definition of the CIDOC Conceptual Reference Model. Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. Version 6.0, 2015. Retrieved, January 28, 2015, from http://www.cidoc-crm.org/docs/cidoc_crm_version_4.2.pdf.
- DCMES. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved January 10, 2007, from <http://www.dublincore.org/documents/1998/09/dces/>.
- Doerr, M. Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM. Technical Report 274, ICS-FORTH, Greece, 2000.
- Gruber, T. (2001). What is an Ontology?. Stanford University. 2001. Retrieved, February 10, 2015, from <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>.

Nilsson, M. (2010) From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization. 2010. Doctoral thesis - KTH School of Computer Science and Communication, Stockholm, Sweden, 2010.

LD4PE: A Competency-Based Framework for DCMI's Professional Education and Training Agenda

Thomas Baker
Sungkyunkwan University,
South Korea
tom@tombaker.org

Michael D. Crandall
University of Washington,
USA
mikecran@uw.edu

Stuart A. Sutton
University of Washington,
USA
sasutton@uw.edu

Marcia Lei Zeng
Kent State University, USA
mzeng@kent.edu

Keywords: Linked Data; LD4PE; competency frameworks; ASN-DF; competency-based discovery; learning resources

1. LD4PE Project Description

This poster reports on early progress of the Linked Data for Professional Education (LD4PE) project to develop a competency-based referatory of learning resources for teaching and learning Linked Data practices in design, implementation, and management. Funded by a grant from the U.S. Institute for Museums and Library Services (IMLS), the project builds on a 2011 IMLS Planning Grant that explored the feasibility and form of an online Linked Data *Exploratorium* of learning resource mapped to a *Competency Index for Linked Data (Index)* that would provide students, professionals, and instructors in the GLAM fields (galleries, libraries, archives, and museums) with structured access to learning resources about Linked Data technology. Learning resources elucidating specific professional competencies are being described and indexed according to knowledge, skills, and habits of mind they embody and are accordingly clustered for discovery and exploration. The benefits of machine actionable data denoting expected competencies have been widely recognized (Ward & Nickolas, 2010). While the *Exploratorium* environment will focus on supporting development of professional competencies related to Linked Data, the project resources, toolkit, *Index*, and website developed for the project will also exemplify those principles and practices.

The *Index* at the center of the project will be a cohesive, stakeholder-developed set of RDF-modeled assertions defining competencies, knowledge, and skills needed for using Linked Data in the GLAM environment. The *Index* will be published as Linked Data in both human-readable and machine-actionable forms using Resource Description Framework (RDF). Individual competency assertions in the *Index* will be assigned globally unique Web identifiers (URIs) to assist in aggregating learning resources about Linked Data practice from across the Web for discovery and exploration by both learners and instructors.

In addition to providing metadata about learning resources, the *Exploratorium* will supply a toolkit, adapted from existing tools and services, that enables creation and subsequent discovery of: (1) RDF metadata describing learning resources from across the Web; and (2) learning maps expressing curricular structures or personal learning trajectories superimposed over the competency framework. These learning maps, also published in RDF, will provide learners and instructors with cognitive scaffolding for approaching the topic of Linked Data.

While the *Exploratorium* is intended to harvest and aggregate learning resource descriptions created by others in RDF and RDFa, the LD4PE Project will initially seed the environment with project-generated descriptions to demonstrate the site's service potential and allow for formative assessment and refinement of the competency framework, toolkit, Web environment, and best-practice documentation. This "seeding" will include creation of learning resource descriptions for

existing resources and for resources created by project partners to exemplify the focused, recipe-like resources the *Exploratorium* will feature going forward.

The Exploratorium website will also support social recommendation mechanisms to highlight the best learning resources and their alignment to the Index. The environment will provide built-in broadcast and responsive communication channels for community engagement and continuous feedback. Figure 1 provides a high-level view of the intended architecture.

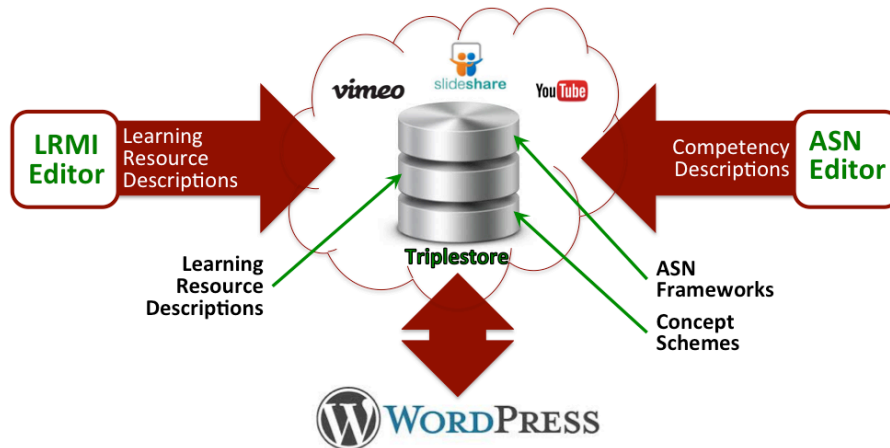


FIG.1 Exploratorium architecture

LD4PE project outputs include: (1) an RDF-modeled *Competency Index for Linked Data* based on the Achievement Standard Network Description Framework (ASN-DF)(Sutton & Golder, 2008); (2) a toolkit to support the generation of RDF metadata; (3) a set of cataloged learning resources, with some developed as exemplars by project partners and others discovered from across the Web; (4) an *Exploratorium* website for the learning resource metadata, toolkit, learning maps, and supporting resources; and (5) best practice documentation available through the *Exploratorium* for all processes, from competency framework development through learning resource development and description to learning map generation.

2. The *Exploratorium* in the Context of DCMI

On successful completion of the LD4PE grant work, the *Exploratorium* will be maintained by the Dublin Core Metadata Initiative (DCMI) as a basic framework for development of its education and training agenda. While focused initially on Linked Data, the lessons learned through development of the *Competency Index for Linked Data* will inform similar competency framework development in other areas of interest to DCMI.

References

- Sutton, Stuart A. and Diny Golder. (2008). Achievement Standards Network (ASN): An application profile for mapping K-12 educational resources to achievement standards. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008, 69-79. Retrieved, April 11, 2015 from <http://dcpapers.dublincore.org/ojs/pubs/article/view/920/916>
- Ward, Nigel and Nick Nicholas. (2010). Benefits of Machine Readable Curricula. Retrieved, April 11, 2015 from http://www.achievementstandards.org/sites/default/files/BenefitsMachineReadableCurricula_Mar10.pdf.

How Should We Teach Metadata? What Comparisons Between Job Ad and Classroom Trends Can Tell Us About Preparing LIS Students

Deborah Maron
University of North Carolina
Chapel Hill, USA
maron@ad.unc.edu

Jacob Hill
University of North Carolina
Chapel Hill, USA
jacobhill.mail@gmail.com

Keywords: metadata; education; pedagogy; job ads; information organization; Linked Data; cataloguing; content analysis; AUTOCAT

1. Context

Metadata is the cornerstone of Galleries, Libraries, Archives and Museums (GLAM) and Digital Humanities (DH) enterprises, and is a fundamental aspect of data management discourse. Information professionals with metadata knowledge are situated as central players in such environments, while those workers lacking such expertise are typically encouraged to acquire it. Metadata literacy, a term defined by Erik Mitchell and used by other scholars, is thus essential for current and nascent information professionals alike (Mitchell 2009).

2. The Problem

Studies detailing the cataloguing and metadata skills required in jobs exist in the literature (Ataman, 2009; Boydston & Leysen, 2014; Chapman, 2007; Hall-Ellis, 2006; Han & Hswe, 2010; Hider, 2006; Millner, 2009; Park & Lu Caimei, 2009; Park, Lu Caimei, & Marion, 2009; Riemer, 2009; Sun Li, 2008; Veve & Feltner-Richert, 2010; Zhu Lihong, 2008). However, the authors hypothesize that the types of knowledge and skills specified in metadata job ads have shifted in the last ~4 years, and yet, there has been no well-publicized content analysis detailing these changes. This lack of research leaves professors of information organization and metadata without a standard for prioritizing the many subjects they could potentially teach in their courses. They are left wondering whether the content they have chosen will adequately prepare their students for the job market.

3. The Study and Anticipated Significance of Findings

In an effort to identify emerging trends in metadata employment and potential deficits in metadata education, the authors extend a study originated by Marcia Zeng, using identical sampling and methodology. Within Zeng's study, a content analysis was performed on five years of AUTOCAT job ads (2007-2012) which were collected manually from the archives and segmented into Excel spreadsheets according to classification (title, skills required, skills desired, etc.). AUTOCAT is a listserv dedicated to issues related to metadata, cataloguing and classification. Because of AUTOCAT's specialization, employers regularly post job ads seeking LIS professionals with the aforementioned skills. Zeng analyzed trends regarding vocabularies, the presence (or absence) of MARC, various metadata standards, Linked Data, programming language requirements and more. Since this is a continuation of an earlier study, it was necessary to limit our data to the AUTOCAT list for the sake of continuity. The authors considered including other sources such as LinkedIn and CODE4LIB but concluded that because these sources are directed at different audiences they were unsuitable for their purposes. The authors are now recording AUTOCAT job competency requirements from October 2012-April 2015. The trends from the cumulative period of 2007-2015 will then be analyzed and the results visually summarized in the accompanying poster.

Upon completion of this initial survey, a follow-up study analyzing terms gleaned from the knowledge organization-related syllabi of LIS programs shall determine if educational trends match hiring trends. The authors argue that the findings will be significant for instructors attempting to align their instruction with needs in the job market, and will complement recent studies on information organization courses and professional development in the classroom (Bibbo & d'Erizans, 2013; Joudrey & McGinnis, 2014). Finally, findings will be significant to the authors of this study as they further develop their own knowledge organization syllabi and digital tools for metadata pedagogy. The authors acknowledge that conducting similar work in other fields would aid their interpretation of this data; however, an extended study it is beyond the scope of this current work. The intention is to illuminate metadata needs for GLAM and DH only at this time.

Acknowledgements

The authors would like to thank Marcia Zeng for her assistance and permission to update her study. They would also like to sincerely thank Adrian Ogletree and Jane Greenberg of Drexel University's Metadata Research Center for their support in presenting the accompanying poster at DCMI 2015.

References

- Ataman, B. K. (2009). Requirements for information professionals in a digital environment: some thoughts. *Program: Electronic Library & Information Systems*, 43(2), 215–228. <http://doi.org/10.1108/00330330910954415>
- Bibbo, T., tamatha_bibbo@asl.org, & d'Erizans, R., roberto_derizans@asl.org. (2013, December 11). The Future of the Librarian as a Metadata Specialist. *Library Media Connection*, 32(3), 26–28.
- Boydston, J. M. K. ., jboydsto@iastate.edu, & Leysen, J. M. . (2014). ARL Cataloger Librarian Roles and Responsibilities Now and In the Future. *Cataloging & Classification Quarterly*, 52(2), 229–250. <http://doi.org/10.1080/01639374.2013.859199>
- Chapman, J. W. (2007). The Roles of the Metadata Librarian in a Research Library. *Library Resources & Technical Services*, 51(4), 279–285.
- Hall-Ellis, S. D. (2006). Cataloging Electronic Resources and Metadata: Employers' Expectations as Reflected in American Libraries and AutoCAT, 2000-2005. *Journal of Education for Library & Information Science*, 47(1), 38–51. <http://doi.org/10.2307/40324336>
- Han, M.-J., & Hswe, P. (2010). The Evolving Role of the Metadata Librarian: Competencies Found in Job Descriptions. *Library Resources & Technical Services*, 54(3), 129–141.
- Hider, P. (2006). A Survey of Continuing Professional Development Activities and Attitudes Amongst Catalogers. *Cataloging & Classification Quarterly*, 42(2), 35–58. http://doi.org/10.1300/J104v42n02_04
- Joudrey, D. N., & McGinnis, R. (2014). Graduate Education for Information Organization, Cataloging, and Metadata. *Cataloging & Classification Quarterly*, 52(5), 506–550. <http://doi.org/10.1080/01639374.2014.911236>
- Millner, M. (2009). *A Content Analysis of Job Descriptions for the Position of Metadata Librarian at Eleven American Universities*.
- Mitchell, E., (2009). *Metadata literacy [electronic resource] : an analysis of metadata awareness in college students*. University of North Carolina at Chapel Hill, Chapel Hill, N.C.
- Park, J., & Lu Caimei. (2009). Metadata Professionals: Roles and Competencies as Reflected in Job Announcements, 2003-2006. *Cataloging & Classification Quarterly*, 47(2), 145–160. <http://doi.org/10.1080/01639370802575575>
- Park, J., Lu Caimei, & Marion, L. (2009). Cataloging Professionals in the Digital Environment: A Content Analysis of Job Descriptions. *Journal of the American Society for Information Science & Technology*, 60(4), 844–857. <http://doi.org/10.1002/asi.21007>
- Riemer, J. J. (2009, February 1). Copy Cataloging as a Catalyst for New Metadata Roles in Cataloging Units. *Technicalities*, 29(4), 1–11.
- Sun Li. (2008). A metadata manager's role in collaborative projects: The Rutgers University Libraries experience. *Electronic Library*, 26(6), 777–789. <http://doi.org/10.1108/02640470810921574>

- Veve, M., & Feltner-Richert, M. (2010). Integrating Non-MARC Metadata Duties into the Workflow of Traditional Catalogers: A Survey of Trends and Perceptions among Catalogers in Four Discussion Lists. *Technical Services Quarterly*, 27(2), 194–213. <http://doi.org/10.1080/07317130903585477>
- Zeng, M. (2014-2015). Personal correspondence.
- Zhu Lihong. (2008). Head of Cataloging Positions in Academic Libraries: An Analysis of Job Advertisements. *Technical Services Quarterly*, 25(4), 49–70. <http://doi.org/10.1080/07317130802128072>

The Sweetpotato Ontology

Vilma Rocio Hualla
International Potato Center,
Peru
v.hualla@cgiar.org

Reinhard Simon
International Potato Center,
Peru
r.simon@cgiar.org

Robert Mwanga
International Potato Center,
Nairobi
r.mawanga@cgiar.org

Henry Saul Juarez Soto
International Potato Center,
Peru
h.juarez@cgiar.org

Genoveva Rossel Montesinos
International Potato Center,
Peru
g.rossel@cgiar.org

Keywords: breeding; ontologies; sweetpotato; phenotypic; genotypic.

1. Introduction

The sweetpotato ontology is part of a community effort to establish a set standard nomenclature to describe crop development and agronomic traits to facilitate analyzing and sharing of phenotypic and genotypic information. The development and adoption of data standards is vital to the interoperability of sweetpotato data (Simon et. al. 2014). Phenotype ontologies are controlled, hierarchically-related phenotypic descriptions that enable large-scale computation among individuals, populations, and even multiple species (Hoehndorf et al., 2013). The International Potato Center (CIP) is currently pursuing the development of standards for plant phenotyping data in collaboration with other interested groups.

The advantage of ontology is that both humans and software applications can understand a data domain. This will allow the application of numerical or data mining techniques that may help to uncover previously unknown correlations. Building on previous draft versions, here we focus on traits important to breeding.

2. Materials and Methods

Through collaborations under the Generation Challenge Program, compatibility data is consolidated by ontologies. Descriptors used in morphology were taken from Huaman (2001). Descriptors used in evaluations were previously standardized (Grüneberg et al., 2009). Additionally, we used descriptors from the Catalogue of Orange-fleshed sweetpotato varieties for Sub-Saharan Africa (Kapinga et al., 2010).

We used the Crop Trait Dictionary Upload Template Version 4 to update the information in the web crop ontology. Terms in ontology were organized in the form of a tree. The nodes of the tree represent entities at greater or lesser levels of detail (Smith, 2004). The branches connecting the nodes represent the relation between two entities (ej. radicle emergence stage is a child of the parent term germination stage). Individual stages of a scale are then parts that can be related to the whole by their order of appearance during plant growth. Each term carries an unique identifier and strictly specified relationships between the terms allow systematic ordering of data within a database, this in turn improves input and retrieval of information (Bard and Rhee, 2004; Harris et al., 2004).

3. Results and Observations

The sweetpotato ontology currently describes 109 traits (Table 1). These include: morphological (28), agronomical performance (28), biochemical (23), reaction to biotic stress (7) and quality traits (23). These traits describe phenotypic variability for characteristics needed for crop improvement. We anticipate further refinements and cross-checks.

TABLE1. Frequency of sweetpotato ontology

Variable	Frequency absolute	Frequency relative
Agronomical traits	28	0.26
Morphological traits	28	0.26
Biochemical traits	23	0.21
Quality traits	23	0.21
Biotic stress	7	0.06
Total	109	1

The ontologies are managed in spreadsheets for ease of transfer from existing data dictionaries. The first version of sweetpotato ontology can be found online.¹

The screenshot displays the Sweetpotato trait ontology interface. On the left, a hierarchical tree shows the structure of the ontology, with 'Morphological traits' expanded to show various sub-traits like 'Abaxial Leaf Vein Pigmentation', 'Distribution of Secondary Flesh color', 'Flower color', 'General Outline of the Leaf', 'Ground Cover', 'Immature Leaf Color', 'Intensity of Predominant Skin color', 'Latex Production in Storage Roots', 'Leaf Lobe Number', 'Leaf Lobes Type', 'Mature Leaf Color', 'Mature Leaf Size', 'Oxidation in Storage Roots', 'Petiole Pigmentation', and 'Plant Type'. The 'Flower color' trait is selected, showing its sub-traits: 'Observation of flower color' (method_of), 'General Outline of the Leaf' (is_a), 'Ground Cover' (is_a), 'Immature Leaf Color' (is_a), 'Observation of immature leaf color' (method_of), 'Intensity of Predominant Skin color' (is_a), 'Latex Production in Storage Roots' (is_a), 'Leaf Lobe Number' (is_a), 'Leaf Lobes Type' (is_a), 'Mature Leaf Color' (is_a), 'Mature Leaf Size' (is_a), 'Oxidation in Storage Roots' (is_a), 'Petiole Pigmentation' (is_a), and 'Plant Type' (is_a). On the right, the detailed view of the 'Observation of flower color' trait is shown, including its identifier (CO_331:0000047), description (Visual categorization), name of method (Observation of flower color), creation date (Mon Jun 16 08:40:50 UTC 2014), and bibliographic reference (CIP, AVRDC, IBPGR. 1991. Descriptors for Sweet Potato. Huamán, Z., editor. International Board for Plant Genetic Resources, Rome, Italy).

FIG 1: Sweetpotato trait ontology structure trait group relationships

4. Conclusions

The goal of ontologies is to construct a set of clearly defined vocabularies that can be used to construct queries between different crops linking the phenotypic and genetic data useful for integrated breeding through data annotation. In addition, the ontologies should enhance future

¹ http://www.croponontology.org/ontology/CO_331/Sweet%20potato

efforts to explore the relationships among phenotypic similarity, gene function, and sequence similarity in plants, and to make genotype-to-phenotype predictions relevant to plant biology, crop improvement, and potentially even human health.

The use of ontological methods to organize biological knowledge is an active area of research and development. The definition of a set of common terms in sweetpotato is contributing in the development of the BioMart database. The datamart of sweetpotato is organized using the sweetpotato ontology.

Acknowledgements

Authors thanks financial support of Crop ontology community of practice of Bioversity International and Generation Challenge Program (Project Code CM-B7001).

References

- Bard J.B., Rhee S.Y. (2004). Ontologies in biology: design, applications and future challenges. *Nat Rev Genet* 5: 213–222
- Grüneberg, W. J., Eyzaguirre, R., Espinoza, J., Mwanga, R. O., Andrade, M., Dapaah, H. & Low, J. (2009). Procedures for the Evaluation and Analysis of Sweetpotato Trials.
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. 2004. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res* 32: D258–D261
- Hoehndorf R, Dumontier M, Gkoutos GV. (2013). Evaluation of research in biomedical ontologies. *Brief Bioinform.* 14:696–712.
- Huaman Z., 2001. Descriptores de la Batata. Retrieved March 20, 2015, from http://www.bioversityinternational.org/uploads/tx_news/Descriptors_for_sweet_potato_Descripteurs_pour_la_patate_douce_Descriptores_de_la_batata_263_ES.pdf.
- Kapinga, R., S.Tumwegamire, J. Ndunguru, M.I. Andrade, S. Agili, R.O.M. Mwanga, S. Laurie, and H. Dapaah. (2010). Catalogue of Orange-fleshed sweetpotato varieties for Sub-Saharan Africa. International Potato Center (CIP), Lima, Peru. 40p. Retrieved March 10, 2015, from: <http://sweetpotatoknowledge.org/germplasm/breeding/availableaterial/OFSP%20Catalogue%20final%20from%20website.pdf#>.
- Oellrich, A., Walls, R., Cannon, S., Cooper, L., Gardiner, J., Gkoutos, G., Harper, L., Hoehndorf, R., Jaiswal, P., Kalberer, Lloyd, J., Meinke, D., Menda, M.; Moore, L.; Nelson, R.; Pujar, A.; Lawrence, K.; Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods* 2015, 11:10 doi:10.1186/s13007-015-0053.
- Simon, R, Hualla, V.; Eyzaguirre, R.; Cordova, R.; Mwanga, R.; Rossel, G.; and Gruneberg, W. (2014). Progress in developing sweetpotato ontology for breeders. Poster present at Workshop on Crop Ontology and Phenotyping Data Interoperability. Conference, Montpellier, France, 31 March to 4 April, 2014.
- Smith, B. (2004). Beyond Concepts: Ontology as Reality Representation. In AVaL Vieu, ed, International Conference on Formal Ontology and Information Systems. Proceedings of FOIS 2004, Turin, Italy.

Advancing Materials Science Semantic Metadata via HIVE

Yue Zhang
Metadata Research Center
Drexel University, USA
yue.zhang@drexel.edu

Jane Greenberg
Metadata Research Center
Drexel University, USA
janeg@drexel.com

Adrian Ogletree
Metadata Research Center
Drexel University, USA
aogletree@drexel.com

Garritt Tucker
Materials Science and
Engineering Department
Drexel University, USA
gtucker@coe.drexel.edu

Keywords: metadata; materials science; metal; SKOS; HIVE; automatic indexing

This poster reports on the process and the initial results of an ontology for metals developed in Simple Knowledge Organization System (SKOS), a World Wide Web Consortium (W3C) standard, and integrated into the HIVE technology.

1. Materials Science: A Need for Semantic Ontologies

Metadata challenges in the materials science community have surfaced due to national and international data sharing policies and the Material Genome Initiative (NSTC, 2014). Among the more pressing challenges is the need to develop semantic ontologies, given their capacity to support information retrieval and discovery, interoperability, and linking of related resources. Researchers engaged in the Materials Science Metadata Infrastructure Initiative (M²I²) are addressing this need by working with the Helping Interdisciplinary Vocabulary Engineering (HIVE) technology in the area of metals.

Materials science is an interdisciplinary field that is advancing the discovery of new materials and enhancing existing materials. Like many interdisciplinary fields, there is a challenge in developing a single semantic ontology due to the breadth of topics that comprise the field. Materials science spans chemistry, engineering, mathematics, and physics, among other disciplines and sub-disciplines. A more productive way to pursue this challenge is to integrate domain-specific vocabularies dynamically when indexing resources. We are exploring this approach at the Metadata Research Center, Drexel University, as part of the M²I² project, working, initially, in the sub-domain of metals.

2. HIVE—Helping Interdisciplinary Vocabulary Engineering

Helping Interdisciplinary Vocabulary Engineering (HIVE) is an automatic Linked Open Data (LOD) technology that integrates interdisciplinary semantic ontologies encoded with the Simple Knowledge Organization System (SKOS), a World Wide Web Consortium (W3C) standard. The integration is dynamic and takes place during an indexing operation. An overview of HIVE's architecture is found on the HIVE Wiki (https://cci.drexel.edu/hivewiki/index.php/Main_Page) and the Code is in GitHub (<https://github.com/MetadataResearchCenter/hive-mrc>).

3. The Metals Ontology: Methodology

The original corpus of terms for the metals ontology was extracted from a set of Wikipedia pages addressing the topic of metal as a material. We identified and defined terms from this set of pages and established their conceptual relationships by their hyperlinks and categories in Wikipedia. The vocabulary was automatically transformed to the SKOS standard by script,

written in C++. The metals ontology includes 44 concepts. The ontology can be described as a high-level or general controlled vocabulary, and is useful for indexing or topical representation in a digital library or repository.

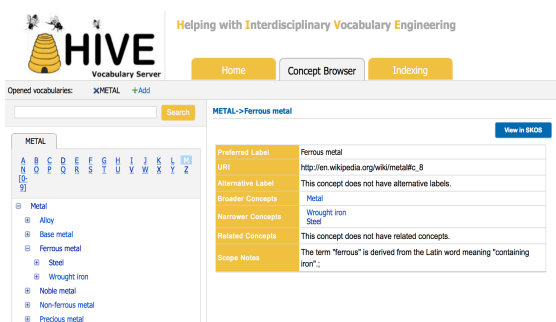


FIG. 1: Ontologies of metals (partial)

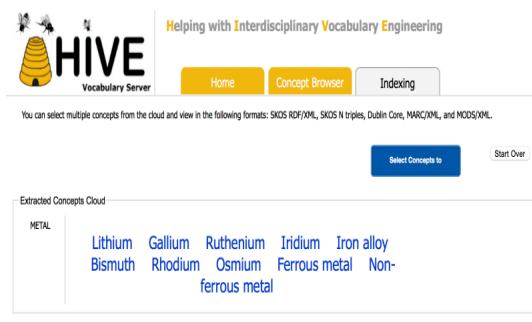


FIG. 2: Sample indexing in materials generated by HIVE

The metals ontology was uploaded to the HIVE demonstration site (<http://hive.cci.drexel.edu:8080/home.html>). The concept browser provides access to the ontology terms via search, and allows access via an alphabetical list (Figure 1). Each concept has narrower, broader, and related terms. For example, the term base metal has a broader term metal, and has narrower terms including Lead, Iron, Nickel, Cooper and Zinc. The indexing operation allows a user to invoke an automatic sequence by uploading a document (e.g. txt, docx, pdf) or entering a URI, for a digital resource, and then selecting the metal ontologies to add. The text is parsed for meaning and matched against the metals ontology. Figure 2 presents the output from running an automatic indexing of the Wikipedia page on metals (<https://en.wikipedia.org/wiki/Metal>) through HIVE. A user can then select appropriate ontology terms from the output for indexing the resource.

4. Status and Next Steps

The initial focus has been to develop a basic and high-level metals ontology for HIVE. A current focus is to enhance HIVE's indexing with the metals ontology, via machine learning algorithms such as KEA++ and MAUI (Frank et al., 1999; Witten et al., 1999). We are working with a group of selected articles and keywords assigned by domain experts (the gold standard) to train HIVE in materials science. We are also working more specifically in the area of naoncrystalline metadata to develop an approach for engaging the scientists in ontology development (Greenberg et al, 2015). Our goal is to further develop semantic ontologies from other sub-domains of Materials Science as we grow HIVE; and as part of this work we will continue to investigate users' preferences for ontologies and functionalities of HIVE.

Acknowledgements

We would like to acknowledge the U.S. National Science Foundation, #OCI-0830944.

References

- Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction.
- Greenberg, J., Zhang, Y., Ogletree, A., Tucker, G. (2015. In press) Threshold Determination and Engaging Materials Scientists in Ontology Design. 9th Metadata and Semantics Research Conference. Manchester, UK, September 9-11, 2015.

- National Science and Technology Council (NSTC). (2014). Materials Genome Initiative: Strategic Plan. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf
- Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999, August). KEA: Practical automatic keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries (pp. 254-255). ACM.

Study of Adhesion between Dublin Core and Marc: Reviewing the Interoperability between UNESP and the National Library

Elaine Parra Affonso
Unesp, Marília
elaineaffonso@marilia.
unesp.br

Elizabete Cristina de Souza
de Aguiar Monteiro
Unesp, Marília
beteaguia@yahoo.com.br

Ricardo César Gonçalves
Sant'ana
Unesp, Tupã
ricardosantana@marilia.
unesp.br

Keywords: metadata; Dublin Core; Marc 21; bibliographic records; interoperability

Abstract

This poster presents a study of interoperability between the National Library and the libraries of the UNESP, in order to identify the adhesion's degree between the MARC 21 standards and the Dublin Core fields present in import bibliographic records from these libraries. Quotations of 50 or more words should be set off as a separate text block using the {Quotation} template element.

1. Introduction

In the description of libraries bibliographic records of, the use of metadata is configured as a key element for parameterization, providing interoperability between databases and systems, enabling better documentary representation and therefore recovery of bibliographic records.

Metadata constitute also a fundamental element in the descriptive treatment process information because they reflect the conjunction of technological and representation needed for new types of resources and information environments, contributing to efficiency of recovery processes in digital environments (Alves, 2010).

The term metadata has different settings according to the area and the application context and or analysis. "Metadata understood as information on data, are intended to document and organize in a structured way, data sets [...]" standardizing them and thus, minimizing rework and facilitating the maintenance of these data (Smith, Costa Santos, 2004, p. 96). The data structure may be obtained by means of sets of pre-defined elements identified by labels (tags) and their respective attributes. A well structured and usage pattern recognized internationally information provides greater data reliability (Rosetto, 2003).

The Dublin Core metadata standard was created in 1995 and has 15 basic elements for describing a variety of features in different information environments Web and used in numerous implementations (Dublin Core Metadata Initiative, 2011).

The MARC 21 format has a structure that allows the construction of bibliographic records so that this represents a variety of types of information resources, facilitates the retrieval of specific information systems resources and promotes the exchange and sharing of bibliographic records between libraries (Alves, 2011).

The standard Dublin Core metadata, as this is responsible for the description of information resources on the Web was chosen, with basic elements of description of the features, and the standard of Machine-Readable Cataloguing metadata (MARC 21), responsible for the description of information resources in bibliographic domain, which consists of a highly structured metadata standard, with complex elements, which uses specific standards and codes for description of resources.

Interoperability obtained in the correlation between different standards is a pressing need, but always occur losses and the possibility of noise in the composition between fields can impair this

process. Libraries configure itself as one of the places where this need for interoperability is present, however, there are situations where the exchange of information takes place between systems that operate with different standards such as MARC 21. The MARC 21 format is specific to area librarianship and the most used in the bibliographic domain (Alves, 2011).

In this context, it is relevant to reflect on the following question: Is there a grip on the match of the fields when the importation of bibliographic records between MARC 21 and Dublin Core standards performed? In this research is being prepared a study that uses the systems and interoperability between the National Library and the libraries of the UNESP.

The network of libraries of the Universidade Estadual Paulista - UNESP besides being part of the consortium with the Fundação Getúlio Vargas is to import the records of the National Library (BN) of Brazil. These records are described in the Dublin Core metadata standard and MARC 21. When working with the cooperative cataloging importing and providing their bibliographic records, promotes interoperability between systems.

The proposition of this study is to assess the degree of adherence in correspondence of the fields present on the importation of bibliographic records between these libraries.

The focus of this study will be linked to the area of information science in the context of information resources described using metadata standards Dublin Core and Marc.

As methodology, the work is feature exploratory and is based on comparison that uses as an analytical tool the bibliographic records imported from the National Library for the UNESP Libraries, identifying which data is used, corrected or discarded and thereby verifying the correlation between standard fields MARC 21 and Dublin Core standard in this particular context.

2. Partial considerations

It is hoped that this study identify, in communication between libraries and even for similar situations, the degree of adhesion between the MARC 21 and Dublin Core standards in a real situation application, thereby providing elements for integration layers can be improved and or parameterized so that one can increase interoperability.

References

- Alves, R. C. V. Metadados como elementos do processo de catalogação. 2010. 132 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.
- Dublin Core Metadata Initiative, 2011. User guide. 2011. Disponível em: <http://wiki.dublincore.org/index.php/User_Guide>. Acesso em: 11 abr. 2015.
- Rosetto, M. Metadados e formatos de metadados em sistemas de informação: caracterização e definição. 2003. 95 f. Dissertação (Mestrado em Ciências da Comunicação)–Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2003.
- Siqueira, M. A.; Costa Santos, P. L. V. A. A versão XML do MARC21 e as formas de representação descritiva na Ciência da Informação. In: VIDOTTI, S. A. B. G. (org). Tecnologia e conteúdos informacionais: Abordagens teóricas e práticas. São Paulo: Polis, 2004, p. 95 – 111.

Bringing a Small Archival Collection to Life on the Web: Remembering the Real Winnie

Sally Wilson
Ryerson University,
Canada
swilson@ryerson.ca

Marina Morgan
Ryerson University,
Canada
marina.morgan@gmail.com

Keywords: schema; metadata mapping; cataloguing; digital collections; digitization; Dublin Core; World War I; Great War; bear; Winnie; Winnie-the-Pooh.

1. Abstract

The purpose of this poster is to provide insight into the processes involved in creating an interdisciplinary online exhibition focused on a unique chapter of Canadian history from World War I. The exhibition focuses on the Colebourn Family Archive comprising digitized photographs and ephemera of Canadian soldier and veterinarian Harry Colebourn (1887–1947) who purchased a pet bear named Winnie who later became A. A. Milne's inspiration for the classic Winnie-the-Pooh children's book series.

2. Introduction

Remembering the Real Winnie: The World's Most Famous Bear Turns 100 is a collaborative, interdisciplinary project that focuses on a unique chapter of Canadian history from WWI. It is based on the Colebourn Family Archives, a collection of photographs, diaries, images, books, and objects, which has been lent to Ryerson University for the purposes of this project. The online exhibit presents the archival content of the collection along with browsable diaries, dynamic maps, and interactive 3D objects.¹

3. Background

Harry Colebourn was a Canadian veterinarian who, on his train journey from Winnipeg to Valcartier to join the Canadian troops heading to Europe at the beginning of WWI, purchased a bear cub in White River, Ontario, for 20 Canadian dollars. This cub was the mascot for Colebourn's regiment and was eventually donated to the London Zoo when the regiment deployed to France. While at the London Zoo, Winnie, named after Winnipeg where Colebourn lived, became popular with the public in general and with Christopher Robin Milne in particular. Christopher Robin called his teddy bear after Winnie, giving it the name Winnie-the-Pooh. This bear was the genesis of the Winnie-the-Pooh storybooks by Christopher Robin's father, A.A. Milne.

4. Research Significance

Through the creation of this online collection, this project has successfully brought various expertise together to explore innovative pedagogical practices. It provided the opportunity for students to gain experience in their fields of study and for librarians to contribute their expertise in designing an online environment for the preservation and analysis of photographs, texts and historical artefacts. This collaborative effort involved cataloguing, metadata mapping, digitization, and website design. The scholarly, online collection promotes research, teaching and learning and demonstrates the value of including the library in this type of collaborative project.

¹ <http://therealwinnie.ryerson.ca/collection/>

5. Tools Used to Create the Digital Collection

5.1. Omeka

To create this digital collection we used Omeka, a free, flexible, and open source web-publishing platform. Omeka allows strong and flexible approaches to metadata representation, easy plug-in deployment, custom implementation of item types, and the addition of the full set of Dublin Core properties to the existing Dublin Core element set, including element refinements (Kucsma, Reiss, and Sidman). With the inclusion of community contributed plugins, you can import a wide variety of data in different formats, create maps, collect information from users, add tags, create timelines and more. For the Winnie project we used the CSV import, Dropbox, Extended Dublin Core, Internet Archive Book Reader, Geolocation and Simple Page plugins. We also customized the basic Omeka software to allow for the inclusion and display of 3D objects.

5.2. Flipbooks

Harry Colebourn kept several diaries during WWI. In the online collection these diaries could only be displayed open at selected pages. By using the Internet Archive BookReader plugin in Omeka we made the entire contents of the diaries browsable. Transcriptions were also made of the diary entries so that the content would be fully accessible to search engines both within Omeka and on the web.

5.3. 3D Scanning

Many of the items in the Colebourn Family Archive are three-dimensional objects including Harry Colebourn's vet bag and its contents. Photographic images were provided for these objects but the 2-dimensional representations are limited in how much information they can convey. A collaboration with the Department of Architectural Sciences at Ryerson enabled experimentation with scanning some of the contents of the vet bag and the bag itself. Since Omeka doesn't deal with 3D files natively, we loaded the files to SketchFab, a YouTube like service for 3D scans and embedded links to them from within Omeka. Users of the website are able to view the 3D objects in their browser and manipulate the object to see all sides.

5.4. Geolocation and Mapping

Omeka has a geolocation plugin, which was used to geolocate photographs with known locations on a map. To supplement this geolocation feature, an external mapping resource was used to create enhanced customized maps. Location information that Harry Colebourn recorded in his diaries during WWI was geocoded and added to the exhibit, along with information about leaves, duration of time spent in each location, types of visits, etc.

6. Metadata Implementation

Data for the project was supplied to the Library in the form of an Excel spreadsheet with multiple workbooks. Extensive corrections of the initial data were done to ensure both compatibility with Omeka software and the appropriate metadata standard, and consistency across the collection.

For this project we used Dublin Core metadata, the most widely adopted metadata standard that offers users the greatest flexibility. This descriptive metadata standard uses broad and generic elements that facilitate the discovery of resources, and provide contextual information useful in the understanding of the resources. Dublin Core provided controlled and structured descriptions of the resources through access points such as title, author, date, location, description and subject. Library of Congress subject headings were added to the metadata as were keyword tags to provide better access both within the collection and to optimize the data for discovery on the web via search engines.

Additionally, separate Dublin Core records were created for each scanned item (photographs, diaries, scrapbooks) some located separately from the resource it describes, others embedded or packaged with it. Since many of the objects were three-dimensional, appropriate descriptive elements needed to be considered. In addition, decisions had to be made about which metadata elements should be displayed and how many were required to make the best use of the metadata within the website and more broadly on the web.

7. Challenges

We experienced considerable challenges with organization of the collection and the descriptive metadata. Omeka is organized with collections, items and files. We determined that a diary would be an item that worked well for creating flipbooks, but didn't work well for the transcriptions, which were eventually created as separate html pages and not within the item metadata. We also experienced some difficulties with the Excel spreadsheets of metadata as they were created by someone without any Dublin Core knowledge and required clean-up before they could be ingested into Omeka.

Furthermore, our inexperience with 3D scanning and its complexities resulted in our underestimating the amount of time required for this portion of the project. All of the items chosen were highly reflective which was problematic for scanning as the light used to make the readings is reflected from the object. We were able to solve the reflectivity problem by using an aerosol spray to coat the reflective objects. This allowed us to capture accurate readings of the geometry of the object, but considerable post-scanning clean up was required to map the surface materials back onto the scan.

8. Conclusions

This project focused on a unique chapter of Canadian and world history, and was brought to light by students, recent alumni and faculty from across the Ryerson campus, who co-developed this multidisciplinary project. This collection was intended to support the research activities of the students and faculty at Ryerson University as well as means of engaging the outside community. Many challenges were encountered during the course of this project. The use of the Dublin Core metadata standard allowed for a broad description of the resources and provided long-term preservation and access to cultural and communicative memories. Furthermore, the aim of the 3D scanning was to explore additional ways of interacting with the objects to see how 3D scanning could be used in this and future digital humanities projects.

Ultimately the project was well managed and run, but the complexity of working with multiple stakeholders and the changing scope of the website portion of the project resulted in several challenges. The website ensures that the entire collection is available to a much broader community for a longer period of time. The scholarly online collection promotes research, teaching and learning, and met its primary goals of increasing access and discoverability to a unique collection.

References

- Dublin Core Metadata Initiative. (2012). DCMI Metadata Terms. Retrieved from <http://dublincore.org/documents/dcmi-terms/>.
- Felicetti, A., Lorenzini, M. (2011). Metadata and Tools for Integration and Preservation of Cultural Heritage 3d Information. In: Proceedings of the 23rd International CIPA Symposium, Prague, 12-16 September 2011. Retrieved from <http://cipa.icomos.org/fileadmin/template/doc/PRAGUE/051.pdf>
- Kucsma, J., Reiss, K., & Sidman, A. (2010). Using Omeka to build digital collections: The METRO case study. *D-Lib Magazine*, 16(3/4) doi:10.1045/march2010-kucsma.
- Mapping Harry Colebourn (2014). Retrieved from <http://therealwinnie.ryerson.ca/collection/mapping1>.
- Omeka Plugins. (2014). Retrieved from <http://omeka.org/add-ons/plugins/>.

- Remembering the Real Winnie 3D Scanning And The Colebourn Family Archive. (2015). Retrieved from <http://therealwinnie.ryerson.ca/collection/scanning>.
- Remembering the Real Winnie Diaries. (2015). Retrieved from <http://therealwinnie.ryerson.ca/collection/>.
- Remembering the Real Winnie Maps. (2015). Retrieved from <https://therealwinnie.ryerson.ca/collection/maps>.
- Remembering the Real Winnie Photographs. (2015). Retrieved from <https://therealwinnie.ryerson.ca/collection/collections/show/3>.

Metadata for Models Generated by openModeller

Agnei Silva
UNASP, Brazil
agnei.silva@hotmail.com

Cleverton Ferreira Borba
University of Sao Paulo -
USP, Brazil
UNASP - Brazil
cleverton.borba@usp.br

Pedro L. Pizzigatti Corrêa
University of Sao Paulo,
Brazil
pedro.correa@usp.br

Keywords: species distribution modeling; Dublin Core; interoperability; openModeller

1. Metadata for Models Tools

In the last years, the economic development has grown on a large scale and accelerating the destruction of the ecosystem process, increasing the demand for tools and methods to support decision making with regard to biodiversity conservation. According to Berendsohn et al. (2011), one of the most serious “bottlenecks” in the scientific workflows of biodiversity sciences is the need to integrate data from different sources, applications software, and services for analysis, visualization and publication.

The main reasons to use metadata patterns in modeling tools are: allow representation of clearer information, interoperate data between repositories, provide standardized structures, increase data accessibility (Dziekaniak, 2010), preserving information resources and documenting legal aspects of resources (Berendsohn et al., 2011). In this context we can explore the Dublin Core metadata, because this help us standardize the models, generated by species distribution modeling tools.

The Global Biodiversity Information Facility (GBIF) and Biological Collection Access Service (BioCASE) are examples of tools that make use of metadata can cite some of them: ABCD (Access to Biological Collection Data) and also the DwC (Darwin Core) metadata that is used to support information from the portals DNA Bank Network and the GeoCASE (Berendsohn et al., 2011). Now, EDIT Platform supports the export and import of data in the standards (ABCD, DwC and also in the SDD - structured Descriptive Data). Among other tools can also cite openModeller (Munoz et al., 2011), receiving information by GBIF and TAPIR/Darwin Core system, utilizing as metadata standards Darwin Core and ABCD.

2. Dublin Core Application for Models Generated by openModeller Tool

Among the existing tools, openModeller stands out with some advantage over other species distribution modeling tools because it allows different formats of data inputs for occurrence of species, environmental data and parameters for the algorithms, above all, different algorithms, simplifying thus to user/users group to reach your aim without needing to know different platforms and modeling tools. One of the problems of the other current tools of species distribution modeling is that they generate models with their standard independent and it cannot be used in other tools.

The need to use the metadata for models generated by OpenModeller tool, allows the data standardization to other platforms, producing data to be reused in the openModeller, and in future, in other tools.

This poster proposes the use of a Dublin Core metadata standard to present and make available the models generated by the species distribution modeling tool openModeller, in order to facilitate interoperability of the data generated by tool itself or other modelling tools.

3. Interoperability of Data Generated

Interoperability only happens, when a well-defined standard is implemented in the data that will be interoperable.

Using an ontology as a class and the Dublin Core metadata as standard, we can ensure that the export and import of the generated models also interoperate between any openModeller tool or any other tool that makes use of species distribution models. Models data will can be available on the Internet so any user will may have free access to this data to visualization or to any other task. Figure 1 clearly describes the idea described in this post:

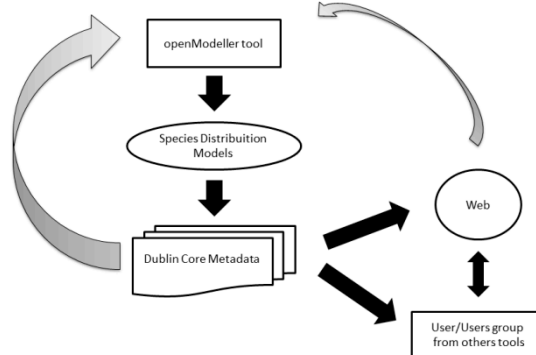


FIG. 1: Interoperability between openModeller tools, other tools and web, using standard Dublin Core Metadata

Therefore, when the openModeller tool generate a package with all information contained in a model, like, algorithm, parameters, occurrence data, etc., could be possible reuse this information to make a new modeling, entering other occurrence data, testing other algorithm and producing or formalizing results.

4. Conclusion

In conclusion, this poster presents a proposal for the reuse of models generated by the species distribution tool, openModeller. It is essential for the reuse of the model, use metadata pattern to ensure biodiversity data interoperability generated by this tools. In this poster we use the Dublin Core metadata for the initial stage of information that need to be reused.

Dublin Core metadata is an important domain to start the standardization of new tools particularly in data generated by species distribution tools that include: algorithm, parameters, climatic packages, biodiversity data, and the model.

Future Research: We suggest the use of new metadata patterns to make more studies case, and if possible, apply this model in other tools of species distribution modeling.

References

- Berendsohn, W. G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K. & Müller, A. (2011) Biodiversity information platforms: From standards to interoperability. *ZooKeys*, v. 150, p. 71-87.
- Dziekaniak, Gisele Vasconcelos. (2010) Mapeamento do uso de padrões de metadados por comunidades científicas.
- Muñoz, M.E.S., Giovanni, R., Siqueira, M.F., Sutton, T., Brewer, P., Scachetti, R.S., Canhos, D.A.L. & Canhos, V.P. (2011) "openModeller: a generic approach to species' potential distribution modelling". *GeoInformatica*.
- Remsen, D., Ko, B., Chavan, B., Raymond, M. (2011) Getting Started, Overview of data publishing in the GBIF Network. GBIF.
- Rodrigues, Fabrício Augusto. (2011) Modelagem da biodiversidade utilizando redes neurais artificiais. Doctoral Thesis presented at University of Sao Paulo.

Evolution of Dublin Core Metadata Standard: An Analysis of the Literature from 1995-2013

Felipe Augusto Arakaki UNESP, Brazil fe.arakaki@marilia.unesp.br	Plácida Leopoldina Ventura Amorim da Costa Santos UNESP, Brazil placida@marilia.unesp.br	Rachel Cristina Vesu Alves UNESP, Brazil rachel@marilia.unesp.br
---	---	---

Keywords: Dublin Core; DC 1995-2013; history of Dublin Core

1. Introduction

Due to the Web development in the beginning of the 90s, new possibilities of sharing information have emerged, especially informational resources. Thus, concerns about the representation and recovery of these resources became the focus of the study of many researchers. In 1995, the Dublin Core metadata standard was proposed to locate and identify any resource in the Web like webpages, textual content and other resources.

Due to its wide scope and the possibility of being used in any context as for example libraries, files, museums, government data, questions arise as to which were the main modifications of Dublin Core since its creation until 2013, and how the development of Dublin Core has occurred in Brazil? The aim was to identify in the literature characteristics that have influenced the development of the Dublin Core standard from reports of its main events and to verify the development of studies about this topic in Brazil.

The work reported is exploratory, qualitative and theoretical research that approaches the evolution and development of Dublin Core standard as the main theme. The work is supported by the São Paulo Research Foundation (FAPESP). The methodology of the work is comprised of a bibliographic survey in the P@rthenon, Capes Portal of journals, Scientific Electronic Library Online (SciELO), Scopus and Web of Science databases. As search strategy, key words like Dublin Core, DCMI, DC, Dublin Core History, Dublin Core Metadata Initiative, Dublin Core events, DCMI Conference, Workshop Dublin Core were used in titles, key words and abstracts, in English, Portuguese and Spanish from 1995 to 2013 in order to identify the reports about Dublin Core Evolution. Later, the focus was on research published in the annals of the Dublin Core Metadata Initiative's (DCMI) *International Conference on Dublin Core and Metadata Applications*, from 1995 to 2013 with the aim of verifying the trends of metadata and Dublin Core research. For the survey in Brazil, databases like the *Biblioteca Digital Brasileira de Teses e Dissertações* (BDTD - Brazilian Digital Library of Thesis and Dissertations), *Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação* (BRAPCI - Referential Database of Journal Articles on Information Science), P@rthenon and SciELO were used. As a search strategy, the term Dublin Core was used in titles, key words and abstracts in the period from 1995 to 2013.

Therefore, the steps for developing the research were the following:

1. Selection of the material, exploratory reading, book report and critical reading;
2. Mapping of researches and characteristics of Dublin Core;
3. Systematization and identification of study characteristics.

2. Results

The research identified that the Dublin Core events can be divided in two periods. The first period corresponds to the Dublin Core Workshop series from 1995 to 2000. The second period started in 2001 and is currently going on through the DCMI conferences. In the first period, eight

events were held, five in North America, two in Europe and one in Oceania. The highlights of the events consisted of: (1) the consolidation of the Dublin Core specifications; (2) the establishment of Dublin Core Metadata Initiative (DCMI) as administrator of specification: and (3) the definition of 15 metadata elements and their qualifiers, among others (Baker, 2012).

In the second period, several themes were discussed like the Dublin Core Abstract Model (DCAM), Dublin Core Application Profile (DCAP), Semantic Web, Linked Data, interoperability among systems and other subjects (Baker, 2012). During this period, the event was held five times in Asia and five times in Europe. North America and Central America held three events. São Paulo State University (UNESP) in Brazil is organizing the first DCMI conference event in South America.

The research followed the consolidation and development of Dublin Core metadata. It work reported here highlights the main points of discussion as identified in the proceedings of the DCMI events such as metadata standardization in different contexts (application profiles), trends of worldwide research on application profiles in domains like museums, libraries, government data, studies related to controlled vocabulary, standardization, harmonization and heterogeneous metadata standard crosswalk. Other highlights are issues related to memory preservation, cultural heritage and digital trusteeship, among others. Through the research reported here, a classification of the works presented in DCMI events is presented according to the main themes studied:

1. Application Profile;
2. Languages;
3. Metadata Mapping;
4. Interference from users in the development of metadata;
5. Sharing and recovering systems information;
6. Thematic treatment;
7. Structure;
8. Bibliometrics;
9. Digital trusteeship; and
10. Web.

In relation to the development of research on the Dublin Core metadata standard in Brazil (Arakaki, Santos & Alves, 2015), 16 articles were identified, all of them in the Information Science area. The results showed research generally linked to companies and institutes of research—for example the Brazilian Agricultural Research Corporation (Embrapa) and the Instituto de Matemática Pura e Aplicada (IMPA). Concerning the thesis and dissertations, five dissertations on Information Science, eight dissertations and two thesis on Computer Science and similar areas were identified.

Among the authors who developed research in Brazil and presented works in DCMI Conferences were Maria E. Catarino from Universidade Estadual de Londrina (UEL)/Brazil, Ana Alice R. P. Baptista from Universidade do Minho/Portugal and the researcher Lucas Vegi from Universidade de Viçosa/Brazil.

3. Considerations

The research identified the evolution and characteristics of Dublin Core metadata. The research survey on Dublin Core standard in Brazil identified three researchers who had their works published in the DCMI International Conference on Dublin Core and Metadata Applications Annals.

Limitations of the work reported here include: (1) surveying the literature in only three languages, as well as the databases that supports the Brazilian literature and the English language; and (2) the classification covered a limited span of the DCMI events (2000-2013). As future work, we anticipate: (1) the classification of research methodology in articles published in

scientific journals—aiming at mapping the studies in journals about Dublin Core metadata standard; and (2) making the data available in the digital environment so that the metadata community have can full access.

Acknowledgements

The authors acknowledge São Paulo Research Foundation (FAPESP) for the financial support (process FAPESP: 12/14274-2) and the members of the New Technologies in Information Research Group (GPNTI) for the discussions during the development of the research.

References

- Arakaki, Felipe A.; Plácida L. V. A. C. Santos & Rachel C. V. Alves. (2015). Panorama das pesquisas sobre o padrão de metadados Dublin Core no Brasil (Overview of research on the pattern of Dublin Core Metadata Standard in Brazil). *Revista ACB*, 20(1), pp. 86-97. Retrieved, June 10, 2015, from <http://revista.acbsc.org.br/racb/article/view/983>
- Baker, Thomas. (2012). Libraries, languages of description, and linked data: a Dublin Core perspective, *Library Hi Tech*, 30(1), 116 - 133. Retrieved, June 10, 2015, from <http://www.emeraldinsight.com/doi/abs/10.1108/07378831211213256>

Adopting the Dublin Core Standard for Describing Open Scientific Data: The e-Quilt Prototype Experiment

Adriana Carla S. de
Oliveira
University of Knoxville,
United States of America
adrianacarla.a@gmail.com

Guilherme Ataíde Dias
Federal University of
Paraíba, Brazil
guilherme@dcf.ccsa.ufpb.br

Renata Lemos dos
Anjos
Federal University of
Paraíba, Brazil
renatalemosdosanjos@gmail.com

Virgínia M. de Souza
Federal University of Paraíba, Brazil
virginiamirandadesouza@gmail.com

Pedro Luiz P. Corrêa
University of São Paulo, Brazil
pedro.correa@poli.usp.br

Keywords: open science; fourth paradigm; data life cycle; dublin core standard

1. The Fourth Paradigm and Open Data

The state of the art in scientific communication is centered on the fourth paradigm. Essentially it brings the open science, open scientific data and the management, sharing, aggregation, curation, preservation and scientific cooperation for the use and reuse of scientific research. We are in the era of intensive data. Hey apud Specht (2015) says:

This is one of the greatest motivations for the re-use of existing data for knowledge creation. With the advancement of technology in capturing and processing data, we have reached the fourth paradigm of data-intensive science and communication, where collaboration between different domain skill sets is required to successfully conduct meta-analysis. (Hey apud Specht, 2015).

Intensive data in the fourth paradigm reinforces the need to improve the skills and to adopt technologies, collaborative tools and methodologies in the context of open science.

Open-data has created an unprecedented opportunity with new challenges for ecosystem scientists. Skills in data management are essential to acquire, manage, publish, access and re-use data. These skills span many disciplines and require trans-disciplinary collaboration. (Specht et al., 2015, p.1)

The ongoing research relies on the data life cycle model and fourth paradigm. The data life cycle adopted for the stages development of the e-Quilt Prototype experiment is the Data Lifecycle developed by the DataONE initiative. This cycle is represented by 8 stages. Tenopir et al. (2011, p. 2) points out the importance of the model,

The collected data are processed through scientific data management and following the data lifecycle model. Different elements can be found in a dataset. For describing the dataset, it is necessary the adoption of metadata standards, follow the data lifecycle for its management and ensure their use and reuse in a long-term. In this way, “the data lifecycle cannot be considered independently from research lifecycle, as data are an indispensable element of scientific research.

The management of scientific open data is shown in Figure 1.

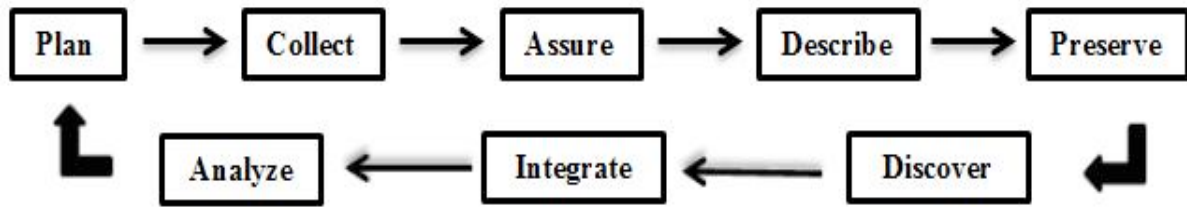


FIG. 1: Management of Scientific Open Data

This phase of the experiment is supported in the *Describe* stage. The data shared in the *e-Quilt Prototype* is the result of the research entitled, Epidemiological Survey on Oral Health, developed by the Department of Social Dentistry, UFPB, held in the cities of Caaporã and João Pessoa, Paraíba, Brazil, in the 2013-2015 period. Primary data collected were shared in the prototype and are being treated according the *Data Lifecycle*.

TABLE 1: Phases of experiment

PHASE 1	Deployment of the Dublin Core Standard to e-Quilt's Prototype Audio Resource	Metadata Description
PHASE 2	Use of the tool <i>Dublin Core Advanced Generation</i>	Automatic metadata code generation
PHASE 3	Audio Resource Metadata Adequacy Analysis	Resource Conformity to the Dublin Core Standard

To describe the metadata contained in the *e-Quilt Prototype* was used elements of the *Dublin Core* standard and the tool *Dublin Core Advanced Generation*.

1.1. Partial Results

The e-Quilt Prototype has the sufficient elements for metadata describing in conformity with international standards. It was verified that the sub-elements and the suggested resources in the Dublin Core standard are likely to be adopted by the metadata associated to the resource analyzed, as shown in Table 2.

TABLE 2: Results of application the audio resource in conformity with Dublin Core standard.

AUDIO RESOURCE – ATTRIBUTE CONFORMITY			
CONFORMITY	DC ELEMENT	CONFORMITY	DC ELEMENT
Yes	20	No	0
Partial	2	N/A	0

The *Identifier* and *Rights* elements associated to the resource presented partial compliance to the standard. The sub-elements DOI and ISBN associated to the *Identifier* element are not used. The analyzed resource is derived from the prototype that has no DOI and the ISBN does not apply to this resource, because it is applied to printed resources. The *Rights* element, presented partial compliance with the License sub-element and was described as unassigned. It was presented in the metadata that the audio resource is in accordance with the Brazilian Copyright Act (LDA - 9.610-1998). This analysis is guided by the adoption of a public license applicable to electronic publications on the international scenario.

The audio resource is derived from the main resource paper, both contained in the ambience of the e-Quilt Prototype. For the audio resource, it was found that it has considerable conformance

to the Dublin Core standard. As for the tool Dublin Core Advanced Generation tool adopted, it was observed that it has limitation concerning the automatic cleaning of characters (symbols, accents, etc.), which should be disposed manually when describing the metadata. Finally, it was analyzed that the description of metadata is a detailed process requiring the adoption of quality criteria and data validation.

References

- DataONE. Data Life Cycle model. Disponivel em: <<https://www.dataone.org>> Accessed: 2015 Mar 15.
- Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, version 1.1. (2014) Available: <http://dublincore.org/documents/dces/>. Accessed 2015 Mar 15.
- Specht, A., et al., Data management challenges in analysis and synthesis in the ecosystem sciences, *Sci Total Environ* (2015), <http://dx.doi.org/10.1016/j.scitotenv.2015.03.092> Accessed: 2015 Mar 05.
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U. et al (June 2011). Data Sharing by Scientists: Practices and Perceptions, *PLoS ONE*. Volume 6, Issue 6. Available: <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101> Accessed: 2015 Mar 06.

Proposal of Application Profile for Digital Images for Libraries, Archives and Museums (DILAM) Conceptual Model

Ana Carolina Simionato
Federal University of São
Carlos (UFSCar),
Brazil
acsimionato@ufscar.br

Plácida L. V. Amorim da
Costa Santos
São Paulo State University
(UNESP), Brazil
placida@marilia.unesp.br

Keywords: DILAM conceptual model; digital image; description of digital image.

1. Introduction

Images have grown on social media and the Web at an exponential level since digital cameras are increasingly available. Several media resources provide digital images, and the archives, libraries and museums need extend possibilities of images use and reuse. Due to the amount of information, the procedures for location and recovery of expressions are difficult tasks to the user. This is an effect of the variety of needed features to describe the digital image.

Therefore, this research focuses on the questioning the conceptual description of the digital image. It is based on the principles of archivology, librarianship and museology. These principles are characterized by the elements of the domain and the structure of the environment used to describe the characteristics of the resource.

The challenge was to represent digital image and specifics elements with an investigative approach. Considering the integrative and divergent features among its descriptive principles of archives, libraries, and museums, the aim is to propose a domain model for digital image resources. The method used in this research is an applied theoretical and qualitative approach in relation to development objectives. In order to clarify the problem of study, the work is also exploratory because the data collection consists of a bibliographic survey at a worldwide level.

2. DILAM conceptual model

The Digital Images for Libraries, Archives and Museums (DILAM) conceptual model was created based on the entity relationship modeling (Simionato, 2015a) that includes the abstractions that these contexts bring to the digital image and the difficulties in creating an image domain. It is important say that DILAM is not a new metadata standard. The DILAM conceptual model was a consequence of the study of conceptual models for specific domains--for example, Functional Requirements for Bibliographic Records (FRBR), Authority (FRAD) and Subject (FRSAD), Conceptual Model for Archival Description (CMAD), Modular Requirements for Records Systems (MoReq) and Conceptual Reference Model (CRM).

The modeling process was based on three steps:

1. The first step derived functional requirements from the parameters of the models studied and descriptive essence of a digital image. The functional requirements to the DILAM are: (a) find or explore the features of image collection, (b) choose the desired pictures between the subjects, using attributes and relationships, (c) recognize the responsibilities of creating a digital image resource, getting the credit, using attributes and relationships, (d) obtain image feature, selected and identified (Simionato, 2015a).
2. The second step consists of choosing the appropriate metadata derived from the crosswalk method (St. Pierre & LaPlant, 1998). Some metadata standards were used, such as: Anglo-American Cataloguing Rules second edition revised (AACR2r), Cataloging Cultural Objects (CCO), Categories for the Description of Works of Art (CDWA), Categories for the Description of Works of Art Lite (CDWA Lite), Describing

Archives content standard (DACS), Dublin Core (DC), Encoded Archival Description (EAD), Graphic Materials, International Standard Archival Description General (ISAD(G)), International Standard Bibliographic Description consolidated edition (ISBD), Resource Description and Access (RDA), Rules for Archival Description (RAD) and SPECTRUM.

3. The last step determined that the qualities of entity relationship modeling could be compatible with entities composed of the FRBR family. Figure 1 shows the DILAM conceptual model, it can also be viewed through the link in references (Simionato, 2015b).

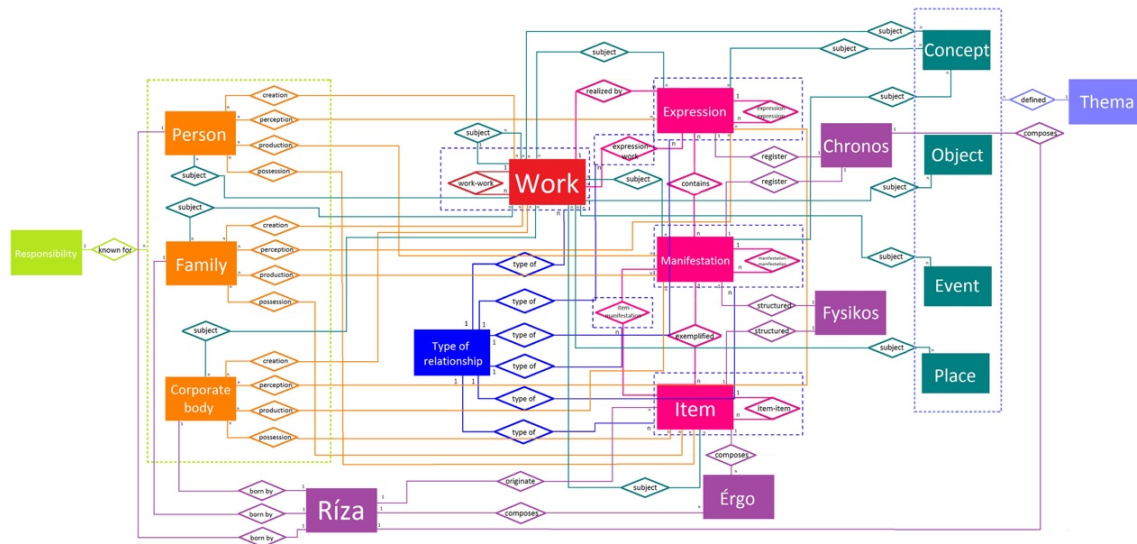


FIG. 1. Digital Images for Libraries, Archives and Museums (DILAM) conceptual model

The entities could also be compatible with the entities that match the integration of contexts. *Chronos*, for instance, is an entity identified in contexts and in the definition of the attributes needed on archives and museums. *Fysikos* is an entity needed for physical properties, as EXIF data. It is a part of the scope of museology in the cautious evaluation of analog image resources and whether there was any damage or other occurrences. *Riza* covers the specific needs for the identification of the origin and provenance. At last, the *Érgo* entity matches the needs that have to be reported, such as the classification, evaluation and curation (Simionato, 2015a).

3. Final considerations

This research brings an approach to the context we live and know, the description in archives, libraries and museums, considering the new needs of linking and integration of data. After all, the sense of this subject among institutions converges and still presents differences. This context is important and it can be collaborative and cooperative with regard to technological advances in the information organization.

Although this research is under development, its results might enable the construction of an application profile based on guidelines for Dublin Core application profiles (Coyle & Baker, 2009). As a result, the domain model DILAM corresponds with the characteristics of the digital image resource. Furthermore, it confirms the collaboration between the descriptive principles of archives, libraries and museums.

References

- Coyle, Karen, Thomas Baker. (2009). Guidelines for Dublin Core application profiles. Retrieved, April 10, 2015, from <http://dublincore.org/documents/profile-guidelines/>

- Simionato, Ana C. (2015a). Modelagem conceitual DILAM: princípios descritivos de arquivos, bibliotecas e museus para o recurso imagético digital (DILAM Conceptual Modeling: principles of archives, libraries and museums for digital image resource). UNESP. Retrieved, March 10, 2015, from <http://hdl.handle.net/11449/123318>
- Simionato, Ana C. (2015b). DILAM Conceptual Model. Retrieved, July 20, 2015, from <https://goo.gl/M7nITN>
- St. Pierre, Margaret, William P. LaPlant (1998). Issues in crosswalking content metadata standards. Baltimore: NISO. Retrieved, March 10, 2015, from http://www.niso.org/publications/white_papers/crosswalk



Best Practice Posters

Joá Archival Description Application Profile: Uma Proposta de Perfil de Aplicação Dublin Core e Encoded Archival Description a Partir da Norma Geral Internacional de Descrição Arquivística [ISAD(G)]

Diana V. B. S. Aleixo
Universidade Estadual de
Londrina, Brasil
dianavbsouto@gmail.com

Maria Elisabete Catarino
Universidade Estadual de
Londrina, Brasil
beteca@uel.br

Ana Alice Baptista
Universidade do Minho,
Portugal
anaalice.baptista@gmail.com

Keywords: General International Standard Archival Description; Resource Description Framework; Encoded Archival Description.; Dublin Core Application Profile.

1 Resumo

Apresenta um perfil de aplicação que faz uso de elementos de metadados dos padrões Dublin Core e Encoded Archival Description, em conformidade com a Norma Geral Internacional de Descrição Arquivística, visando a descrição dos documentos arquivísticos nos moldes do modelo Resource Description Framework. O perfil de aplicação desenvolvido tem como foco auxiliar na organização das informações arquivísticas existentes hoje na web. Desta maneira, este instrumento descreve as unidades arquivísticas que compõem a descrição de um documento em partes, estruturando as informações e relacionando-as outras informações presentes em outras bases de dados por meio de inferências, tendo como objeto um vocabulário específico da área.

2 introdução

Este trabalho propõe um perfil de aplicação que faz uso dos metadados Dublin Core (DC) e Encoded Archival Description (EAD) a partir da Norma ISAD(G), orientado a descrição arquivística nos moldes do modelo Resource Description Framework (RDF). Para alcançar tal objetivo, foram cumpridas as ações relacionadas a: explorar as recomendações elaboradas pelo W3C, Dublin Core Metadata Initiative (DCMI), pelo comitê técnico da EAD, e as Normas de Descrição Arquivística; identificar e analisar as correlações entre os campos de atributos e os campos de descrição existentes na norma ISAD(G) com DC e EAD para adoção do modelo RDF; e compor um perfil de aplicação com os metadados para descrição arquivística no contexto da *Web Semântica*.

3 Desenvolvimento

O perfil de aplicação criado recebeu o nome de Joá Archival Description Application Profile (JADAP), sendo concebido para declarar termos de metadados úteis para a descrição arquivística, já existentes no DC e da EAD, tomando como eixo central a Norma ISAD (G).

A Norma ISAD(G) serviu como guia na proposição dos termos que serviram como descritores do perfil. Assim, foram analisados os 146 termos presentes na EAD e os 15 termos existentes no DC, para então eleger os 15 termos que descrevessem a unidade arquivística de forma a responder à questão que norteou todo este estudo.

O JADAP compreende os termos de metadados necessários para que ocorra uma descrição eficiente e eficaz. Assim, optou-se por nominar a unidade de descrição do recurso a ser descrita como Unit for Archival Description. Desta forma, foram elencadas as seguintes propriedades do DC e da EAD no quadro a seguir:

Quadro 1: ISAD(G): Propriedades do DC e EAD

	ISAD(G)	DC	EAD
ÁREA DE IDENTIFICAÇÃO	3.1.1 Código de referência	Identifier	
	3.1.2 Título	Title	
	3.1.3 Data(s)	Date	
	3.1.4 Nível de descrição	Level	
	3.1.5 Dimensão e suporte	Format Type	
ÁREA DE CONTEXTO ALIÇÃO	3.1.1 Código de referência	Creator	
	3.2.2 História Administrativa/Biografia		Biography for History <bioghist>
	3.2.3 História arquivística	Description	
ÁREA DE CONDIÇÕES DE ACESSO E USO	3.4.1 Condições de acesso		Conditions Governing Access <accessrestrict>
	3.4.2 Condições de reprodução		Conditions Governing Use <userrestrict>
	3.4.3 Idioma	Language	
ÁREA DE FONTES RELACIONADAS	3.5.1 Existência e localização dos originais		Locations of Originals <originalsloc>
	3.5.2 Existência e localização de cópias		Alternative form Available <altformavail>
ÁREA DE INDEXAÇÃO		Subject	

Fonte: Elaborado pela autora com base na Norma ISAD(G) (CONSELHO INTERNACIONAL DE ARQUIVOS, 2001), DCMI Terms (DUBLIN CORE METADATA INITIATIVE, 2005) e no *Encoded Archival Description Best Practices* (ENCODED ARCHIVAL DESCRIPTION BEST PRACTICES WORKING GROUP, 2004).

Assim, o perfil de aplicação proposto pretende auxiliar reunindo termos elencados na Norma ISAD(G) em uma estrutura que possibilite que estes sejam descritos e relacionados a outros termos de igual significado.

A escolha dos termos que compõem o JADAP restringiu-se apenas a Norma ISAD(G), visualiza-se que seja possível unir outros termos presentes nas demais normas de descrição, porém devido a este trabalho ser resultado de um mestrado, constatou-se que não seria possível realizar em tempo hábil a análise de tais documentos, o que suscita a possibilidade de elaboração de novos trabalhos relacionados a esta questão.

Pôde-se constatar durante a elaboração deste estudo, que a produção científica na CI, mais especificamente na Arquivística, relacionada a Perfil de aplicação no Brasil é pequena. Observou-se que o debate no cenário nacional, unindo as temáticas, fica restrito a poucas escolas, e que poucos são os trabalhos que tem seus resultados publicados em revistas ou eventos relacionados as áreas.

Considera-se essencial que o debate iniciado neste estudo de aliar as áreas relacionadas. Para tanto, pretende-se unir a metodologia proposta nesta pesquisa com outras investigações, de maneira a propiciar o desenvolvimento de projetos que venham a contribuir com a organização e recuperação da informação na *web*.

Use of Dublin Core to Increase Public Transparency of Brazilian Senate's Bills Datasets

Fernando de Assis Rodrigues
Universidade Estadual Paulista, Brazil
fernando@elleth.org

Ricardo César Gonçalves Sant'Ana
Universidade Estadual Paulista, Brazil
ricardosantana@marilia.unesp.br

Keywords: Open Government Data; data gathering; Metadata; Dataset; Brazilian Senate

1. Introduction

The transparency of government actions in society is an integral part of discussions about public administration models. These new public management sets seeks to redistribute skills and resources among different within and outside government organizations, allowing an increase in institutional pluralism on public office (Malin, 2006; Sant'Ana & Rodrigues, 2013). One way to strengthen transparency of government actions and to ensure a greater visibility of their activities can be achieved through an expansion of information-sharing environments that among its features provides new information flows between government and society (Rodrigues, Sant'Ana, & Fernalda, 2015). Thus, citizen participation will be extended beyond elections processes and the government will be able to improve their effectiveness and monitoring activities and results of their own actions (Bohman, 2000; Open Government Partnership, 2014).

The Brazilian government establishes citizen rights to claim and access government information and data on a specific legislation called Information Access Law (from Brazilian Portuguese: Lei de Acesso à Informação). This legislation makes mandatory an use of Internet as a dissemination tool, towards to citizens grant access of Brazilian government's data (Brasil, 1988, 2011). Also, it is important that government datasets be machine-readable (Berners-Lee, Hendler, & Lassila, 2001) and available in a way that can be “[...] suitable for use without re-typing or additional treatments to a direct data gathering [...]” (Sant'Ana & Rodrigues, 2013, p. 51) by external agents, independently of an initial format or a specific technology platform.

The goal of this paper is to presents an ongoing study of the applicability of using Dublin Core metadata terms in data retrieval to describe Senate's bills datasets, in order to increase the total amount of describing elements available to government data in gathering process.

The methodology adopted is based on an exploratory analysis of government datasets that were available on the set of Brazilian Senate's websites, on January, 2015. This analysis is divided into three phases: i) search for available bills datasets in Senate websites; ii) explicit metadata elements, already available in retrieved datasets; iii) find available information on these websites, specifically in datasets retrieval area pages, that can be part of a future Dublin Core metadata set of elements.

2. Website and dataset characteristics

The Brazilian Senate has a specific website to share data about its activities called “Portal e-Cidadania – Dados Abertos”. The website have forty five sets of data available, grouped into eight predefined groups. The information resource which contains the government datasets with bills and votes data is called Nominal Bills, and it is located in the 'Plenary Sessions' group. Nominal Bills consists on a set of twelve items, as follows: one description page; nine dump files for download in eXtensible Markup Language (XML) format, containing Nominal Bills data, grouped by year; one hyperlink that redirects users to other Senate's web site for queries, acting like a search interface to citizens; one hyperlink to a web service interface that provides an gateway to an external automated gathering process.

For each item, the website offer an unique page in HyperText Markup Language (HTML) format, divided by sections, as follows (in a top-down order): a) a paragraph with an item description; b) a 'download' button; c) a 'quick information board' with four elements (named: Part of dataset, Last updated, Format and License); d) an Additional Information (in a HTML table format) with two columns (named: Field and Value, not sortable) and nineteen rows with its field names written in lowercase and without blank spaces. All 'quick information board' elements and Additional Information field names are written in English language.

In web service interface for data gathering, it's possible to query Nominal Bills that occurred only by inputting a specific date. For example, to an external application collect these datasets and retrieve all Nominal Bills on a particular month, will be necessary to run 'x' queries, where 'x' represents a total of days in that month.

All queries results are in a XML format, and its elements hierarchy is organized as follows: a root attribute called BillsList (ListaVotacoes), who has two children elements: a) Metadata (Metadados) and b) Bills (Votacoes).

The Metadata element has a fixed number of children elements (three): a) Version (Versao) with its value being the date inputted previously as query parameter; b) ServiceVersion (VersaoServico), an integer with no further description, and; c) DataSetDescription (DescricaoDataSet) with its value for all queries being a fixed text (a sentence about the web service, with information about data updates and two hyperlinks (to a XML file and a XML Schema Definition (XSD) file). On Bills element, each children element represents data from a unique Bill result.

3. Results

When Nominal Bills datasets are gathering by external agents, it is available three metadata elements: Version, ServiceVersion and DataSetDescription. This set remains equal in web service interface and XML dump files. Other two hyperlinks were found in DataSetDescription element value – both redirects users to a Senate's web site error page.

All 'quick information board' elements and all rows in Additional Information section don't have any kind of explanation about its meaning.

4. Conclusion

In Additional Information sections, even rows doesn't have any kind of explanation about its meaning, it is possible that this set had potentially metadata information about Nominal Bills dataset that isn't available yet on retrieved files. For example, a value of 'name' field seems to be a description of dataset content; a value of 'id' field seems to be a unique resource identifier to dataset; etc. That kind of information could be more explored and added as a children element in Metadata element on XML files.

It concludes that on an application of Dublin Core set of descriptive elements on Senate's bills datasets have to observe the following variables: a) a study of meaning of elements found on unique pages; b) development a strategy that takes into account to fill Dublin Core required elements with values already available in Metadata's children elements and in dataset page section's, including an evaluation of adoption external software tools or data conversion algorithms on this process; c) a replacement of existing Metadata's children elements and namespaces.

As future work, it proposes a development of a dataset prototype with an application of the Dublin Core elements in a Nominal Bills XML dump file.

References

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 28–37.
- Bohman, J. (2000). *Public deliberation: Pluralism, complexity, and democracy*. MIT press.
- Brasil. Constituição da República Federativa do Brasil de 1988 (1988).
- Brasil. Lei Nº 12.527, de 18 de Novembro de 2011 (Lei de Acesso à Informação), Pub. L. No. 12.527 (2011).
- Malin, A. M. B. (2006). Gestão da Informação Governamental: em direção a uma metodologia de avaliação. *DataGramaZero*, 7(5).
- Open Government Partnership (Ed.). (2014). Open Government Partnership: four year strategy 2015-2018.
- Rodrigues, F. de A., Sant'Ana, R. C. G., & Ferneda, E. (2015). Análise do processo de recuperação de conjuntos de dados em repositórios governamentais. *InCID: Revista de Ciência da Informação e Documentação*, 6(1), 38–56. <http://doi.org/http://dx.doi.org/10.11606/issn.2178-2075.v6i1p38-56>
- Sant'Ana, R. C. G., & Rodrigues, F. de A. (2013). Visualização de afinidades entre parlamentares mediante dados de votações no Senado Brasileiro. *Informação & Sociedade: estudos*, 23(1), 49–59.

Reutilização de Metadados para o Povoamento de um Repositório Institucional: Procedimentos Aplicados no Repositório Institucional UNESP

Silvana Aparecida Borsetti
Gregorio Vidotti
UNESP – Univ Estadual
Paulista, Brasil
vidotti@reitoria.unesp.br

Flávia Maria Bastos
UNESP – Univ Estadual
Paulista, Brasil
fmbastos@reitoria.unesp.br

Juliano Benedito Ferreira
UNESP – Univ Estadual
Paulista, Brasil
julianoferreira@reitoria.unesp.br

Ana Paula Grisoto
UNESP – Univ Estadual
Paulista, Brasil
grisotoana@reitoria.unesp.br

Fabrcio Silva Assumpção
UNESP – Univ Estadual
Paulista, Brasil
fabricio@reitoria.unesp.br

Renata Eleutério da Silva
UNESP – Univ Estadual
Paulista, Brasil
renata_silva@marilia.unesp.br

Vitor Silvrio Rodrigues
UNESP – Univ Estadual Paulista,
Brasil
vitorsrodrigues@reitoria.unesp.br

Oberdan Luiz May
UNESP – Univ Estadual
Paulista, Brasil
oberdan@reitoria.unesp.br

Palavras-chave: Reutilização de metadados; repositório institucional.

1. Introdução

Os esforços para a implantação de um repositório institucional na Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP) tiveram início em 2013, quando a Universidade optou pelo uso do software DSpace e definiu, a partir do padrão Dublin Core, um conjunto de metadados para alcançar suas necessidades no que diz respeito à produção científica de seus pesquisadores.

A meta inicial do Repositório Institucional UNESP (<http://repositorio.unesp.br>) era incluir os artigos publicados no período de 2008 a 2012 e indexados na Web of Science. Para alcançar essa meta, optou-se pela reutilização dos metadados já existentes na Web of Science para criar registros para importação no DSpace. Após o alcance dessa meta inicial, os metadados de outras bases de dados também foram reutilizados.

Este trabalho apresenta alguns procedimentos para a reutilização de metadados a partir da Web of Science, da SciELO, da Scopus e da Plataforma Lattes na criação de registros para importação no Repositório Institucional UNESP.

2. Coleta

Para coletar os metadados em um formato XML, foram utilizados diferentes procedimentos para cada base de dados: os metadados da Web of Science foram coletados utilizando o *web service* disponibilizado pela Web of Science; os metadados da SciELO foram coletados utilizando um software criado por um membro da equipe do Repositório; os metadados da Scopus foram comprados pela Universidade; e os metadados da Plataforma Lattes foram coletados utilizando outro software criado pela equipe. A ferramenta criada para coletar os metadados da SciELO também coletou os objetos digitais (ou seja, os arquivos PDF).

3. Conversão

As bases de dados utilizadas como fonte de dados possuem seus próprios padrões de metadados, assim, foi necessário mapear os metadados dos quatro padrões para os metadados do perfil de aplicação utilizado no Repositório e converter os registros de modo a obtê-los de acordo com esse perfil de aplicação. Uma vez que os registros foram coletados em XML, foram criadas folhas de estilo com a linguagem Extensible Stylesheet Language for Transformation (XSLT) para realizar a conversão. A conversão ocorreu em dois passos: (1) conversão do arquivo XML original em um arquivo XML de acordo com o perfil de aplicação e (2) conversão do arquivo XML de acordo com o perfil de aplicação em um arquivo CSV. Para os registros da Scopus e da Plataforma Lattes foi necessário um passo adicional antes do primeiro passo para juntar todos os arquivos XML coletados em um único arquivo XML.

4. Verificação

Durante a etapa de verificação, primeiramente foram removidos os registros duplicados. Para isso, os registros foram comparados entre si e com os registros já presentes no Repositório. Para a comparação foram utilizados o DOI e o título e ano de publicação juntos.

Após a remoção das duplicações, os registros foram verificados pela equipe para checar se a Universidade estava mencionada nos dados de afiliação dos autores, para corrigir erros e incluir os dados ausentes. Após essa verificação, foram verificadas também as permissões de acesso (acesso aberto ou acesso restrito) e de arquivamento do objeto digital (se o arquivamento em repositórios institucionais era permitido ou não). Nos casos em que o arquivamento era permitido, uma cópia do objeto digital era salva e nomeada com um ID obtido a partir do registro (o ID da Web of Science, da SciELO ou da Scopus, por exemplo).

Ao final desta etapa, foi executado um programa que distribuiu os registros entre as coleções do Repositório a partir das informações presentes nos metadados de afiliação e de autor. Esse programa, criado pela equipe, incluiu em uma coluna do arquivo CSV o código “handle” das coleções nas quais o registro deveria ser incluído.

5. Importação

O arquivo CSV verificado foi importado no DSpace de modo a inserir os registros no Repositório. Após a importação, um programa desenvolvido pela equipe incluiu cada objeto digital coletado em seu respectivo registro a partir da correspondência entre o nome do objeto digital e o ID presente no registro.

6. Considerações finais

Com a aplicação dos procedimentos apresentados neste trabalho, a reutilização de metadados permitiu o alcance de resultados positivos no Repositório Institucional UNESP: mais de 80 mil registros foram inseridos em cerca de um ano e meio. Esses procedimentos têm como principal característica os mapeamentos entre os padrões de metadados utilizados nas bases de dados (Web of Science, Scopus, SciELO e Plataforma Lattes) e o perfil de aplicação de metadados utilizado no Repositório criado a partir do padrão Dublin Core.

Por fim, com a demonstração dos procedimentos de reutilização dos metadados para importação de itens em um repositório institucional, este trabalho provê contribuições para as instituições que almejam aumentar as coleções de seus repositórios e, conseqüentemente, sua visibilidade acadêmica.

Metadata Extraction and Register for Enterprise Information Architecture in the Brazilian House of Representatives

Mariana Baptista Brandt
Câmara dos Deputados, Brazil
mariana.brandt@camara.leg.br

Keywords: enterprise information architecture; metadata register; metadata identification; House of Representatives (Brazil).

Abstract

This paper presents part of the Enterprise Information Architecture of the Brazilian House of Representatives, which aims to model the information of the strategic business processes and integrate it to their information systems. This procedure intends to be part of the institution's Enterprise Architecture.

The extraction of business metadata is one of the most important parts of the information modeling - the Information Architecture methodology used by the institution. It starts by analyzing the business process using process mapping and modeling with Business Process Management methodology. The information analyst joins the business area and the process management teams in the activity of business mapping and modeling so as to get to know the business and identify the information produced and consumed during the business process.

The information modeling team follows the business mapping and modeling meetings. Depending on the business process, the information modeling team can follow more steps of the process management team to understand the business information. During the business mapping, the team identifies the procedures, activities, information flows and documents needed to the business execution. The business mapping diagram produced is used by the information analyst, who identifies, from each activity, where there is important information input or output. This creates the document "Business Process Information Map", that shows the information and documents used and produced in the business process for business acknowledge and metadata extraction. The process modeling implements adjusts and improvements for the efficacy and efficiency of the business. After it's done, the information modeling team check if there are any changes in the business information and update it if necessary.

The next step consists in the metadata identification from the documents and information of the process analysis. The business area which is responsible for the process provides copies of all the documents identified in the process to the information modeling team, who analyses the documents for metadata extraction. The documents can be from any kind: manuals, reports, IT systems screens, checklists, administrative processes, orientation guides, etc. The metadata extracted from these documents are registered in a repository. For each metadata, are also registered its attributes: description, data steward, access mean, standard entry, format rule, responsible for first entry in the system, business rule, access level and if it's part of an open data dataset. These attributes came from the analysis of the information policies of the institution (Information Content Management Policy,¹ Digital Preservation Policy,² Index Policy,³

¹ Ato da mesa 46/2012. Available at: < <http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-46-16-julho-2012-773828-norma-cd.html>>.

² Ato da mesa 48/2012. Available at: <http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-48-16-julho-2012-773828-norma-cd.html>>.

³ Ato da mesa 50/2012. Available at: < <http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-50-16-julho-2012-773825-norma-cd.html>>.

Publishing Policy,⁴ Information Security Policy,⁵ Freedom of information Law⁶ and it's regulation⁷).

These attributes bring important information about the metadata and can be about the metadata itself or resulting of the relationship between the metadata and the business process. It means that for a determined business process, the metadata can have, for example, an access level different from another business process, among other differences related to the business process. The metadata attributes are: description, data steward and format rule. These are the attributes that identifies the metadata as unique and can't be altered according to the process. If there is a need of changing in these attributes, it should be analyzed if there's a need of creating a new metadata. The other attributes are related to the business process in which the metadata is in.

After identifying the business metadata and representing it with all its attributes, the business area is required to validate it. The information modeling team provides orientations about how the validation must be done and offers help to this task if necessary.

When the metadata are validated they are used for other activities of the enterprise information architecture, such as: information governance, information retrieval requirements, management information needs report and the information architecture diagram. The last one is a diagram that includes metadata, data stewards and information technology systems that are part of the business process. This documentation, in addition to the terminological part developed in parallel by another team (business glossary, taxonomies and thesauri) makes what is called Information Architecture Model. This model must be a guideline to the development of IT solutions and information management tools for the business process.

The register of the metadata and its attributes in a repository makes possible to have a general vision of the institution's information. Besides, it promotes the governance because it shows who the data stewards are. The metadata reuse in more than one business process is another possibility provided by the registration in a repository. The data steward has the authority to decide about the information content characteristics and attributes and its business related metadata.

The metadata mapping and its reuse in different business processes and information systems allows improvements in the information management and information quality, because it avoids non controlled redundancy and inconsistencies. It also highlights the data steward, who must warrant information authenticity, integrity, accuracy and security, and who will be accountable when one of these criteria is not observed.

The enterprise information architecture aims to organize and integrate the business processes information to its IT systems, contributing to the institution's enterprise architecture. The metadata are its more representative element and provide the improvement of information access and information quality.

⁴ Ato da mesa 50/2013. Available at : <<http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-50-16-julho-2012-773825-norma-cd.html>>.

⁵ Ato da mesa 47/2012. Available at: < <http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-47-16-julho-2012-773827-norma-cd.html>>.

⁶ Available at: <<http://www2.camara.leg.br/legin/fed/lei/2011/lei-12527-18-novembro-2011-611802-norma-pl.html>>

⁷ Ato da mesa 45/2012 Available at : < <http://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-45-16-julho-2012-773823-norma-cd.html>>.

Gateway to Oklahoma History Case Study: Structured Data and Metadata Evaluation for Improved Image Resource Findability on the Web

Emily Ann Kolvitz
University of Oklahoma,
USA
kolvitz1@gmail.com

Keywords: structured data; metadata; Semantic Web; online search; information retrieval; image findability

1. Introduction

Image Resource Findability on the World Wide Web is still very much a land-grab. For the Semantic Web to become a reality online businesses and individuals have to get their hands dirty and also come face-to-face with the realization that search engine giants are increasingly becoming the go-to tool for information resource retrieval. “Increasingly, students use Web search engines such as Google to locate information resources rather than seek out library online catalogs or databases of scholarly journal articles” (Lippincott 2013). This puts the search engine giant in a unique position to dictate how the future of search will work on the Web - and therefore, your organization’s future presence (or lack thereof) on the Web. Search Engine Optimization (SEO) techniques change frequently and remain much a mystery to many companies. The one variable in the equation of Web findability that remains a staple is good quality metadata under the hood of the Website. In this case study, a methodology is applied to the Gateway to Oklahoma History’s Website. This study can be generalized to organizations looking to benchmark their own findability maturity on the Web from an image-centric viewpoint.

2. Purpose

Image search and retrieval is a more difficult area than text search and retrieval because accessibility to the image content is largely dependent on the context presented in and around the image resource. The future of Semantic Web technologies relies very much on the idea that organizations are fluent in structured data and have devoted resources to exposing valuable data to the web. The W3C (World Wide Web Consortium) was founded over two decades ago and the widespread adoption of Schema.org and structured data on the Web did not gain traction until the big four search engines (Google, Bing, Yahoo, and Yandex) agreed that a standard was needed to pave the way forward. “On-page markup helps search engines understand the information on web pages and provide richer search results. A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts.” (Schema.org 2014) Many organizations still lag behind on their implementation of any type of structured data. Structured data is only one piece of the findability algorithm. Metadata near content, embedded within content, or listed in the alt text of an html document all tell machines something about the content inside of the record as well. There are no guardians of the Web, ensuring structured data is uniformly applied to all records with equal attention and care and there is no standard, mandated requirement for records on the Web to provide context for image resource findability. Most search engines do not crawl embedded XMP data or the invisible Web, leaving text near images, file names or text in the alt-text in html markup as the only context for image resources. The search algorithms for image retrieval are subject to change frequently (Kritzinger 2013) and additionally, social media sites and organizations strip embedded data from images (Embedded Metadata Manifesto 2014). Embedded metadata provides context and

provenance for image resources. Even with the dramatic adoption of structured data markup utilizing schema.org vocabularies, there still remains metadata opportunities on the Web. Reicks recommends embedded metadata as a strategy for online findability by showcasing examples of applications that parse embedded data into structured data around images on the Web such as PhotoShelter and LicenseStream (2010).

2. Research Methods

The following research question informed this project: What are the types and quality of structured data, XMP, and metadata records available for image resources appearing on the website? Utilizing the Structured Data Linter Tool and Phil Harvey's ExifTool, information was gathered to quantify these research questions. Image records on the Gateway to Oklahoma History's website were investigated for the types, quality and quantity of embedded metadata and structured data.

3. Results

The Gateway to Oklahoma History's Website has a wealth of structured data and metadata pertaining to its image resources. Search queries utilizing structured data markup tags and/or embedded metadata yielded relevant and accurate results during a normal web search, but did not yield relevant and/or accurate image resources during an image search. Descriptive filenames were not used for image resources, which is an important part of image retrieval through web search engines. Adding Schema.org tags to the on-page markup, to accompany the structured data already present is another area for improvement. An interesting finding from this research was that embedded metadata was only found on the largest, original version of the image resource, and never on smaller derivative images. Structured data included in the on-page markup included Open Graph Protocol and Dublin Core. IPTC was the primarily type of embedded metadata present for the image resources.

4. Conclusion

The results and methodology for this research can help GLAM institutions (Galleries, Libraries, Archives & Museums) by bringing awareness to the state of structured data and image resource findability for cultural heritage institutions on the Web. GLAMs must be active in the SEO space, support machine-readable language in the markup of their sites, and utilize Schema.org vocabularies and descriptive filenames for relevancy in search engine results. The Digital Library Federation, which is a program of the Council on Library and Information Resources, concludes that "Getting found means repository objects must be included in the indexes of major search engines because most students and faculty now begin their research with Internet search engines. Digital repositories created by libraries will be largely invisible to users if their contents are not indexed in these search engines" (Digital Library Foundation 2014).

References

- Digital Library Federation. Last accessed October 20, 2014. "SEO for Digital Libraries."
<http://www.diglib.org/community/groups/seo-for-digital-libraries/>
- International Business, Times. 0006. "Bing, Google and Yahoo merge to make search easier with schema.org." International Business Times, April.
- IPTC International Press Telecommunications Council, 2014. "Embedded Metadata Manifesto" Last accessed November 20, 2014. <http://www.embeddedmetadata.org/social-media-test-results.php> (Embedded Metadata Manifesto 2014).
- Kritzing, W. T. "Search Engine Optimization and Pay-per-Click Marketing Strategies." Journal of Organizational Computing and Electronic Commerce, no. 3 (2013): 273-86.

- Lippincott, Joan K. "Net Generation Students and Libraries," EDUCAUSE (2005), accessed November 19, 2014, <http://www.educause.edu/research-and-publications/books/educating-net-generation/net-generation-students-and-libraries>
- Reicks, David. 2010. "Why Embedded Metadata Won't Help Your SEO," Last Updated December 30, 2013. Last Accessed November 23, 2014. <http://www.controlledvocabulary.com/blog/embedded-metadata-wont-help-seo.html>
- Schema.org. 2015. "About Schema.org" Last Updated Unknown. <https://schema.org/docs/faq.html>

Data Harmonisation between National Library Board, National Archives and National Heritage Board of Singapore

Shan Shan Chan	Haliza Jailani
National Library Board of	National Library Board of
Singapore	Singapore
Chan_Shan_Shan@nlb.gov.sg	Haliza_JAILANI@nlb.gov.sg

Keywords: data harmonisation; schema; mapping; crosswalking; single search interface; search engine optimization; controlled vocabularies; name headings; Taxonomy & Thesaurus Editor; OneSearch; cross-linkages

1. Introduction

The NLB Data Harmonisation Project aims to enhance user experience and the discovery of nuggets of resources from the rich collections of the National Library Board (NLB), the National Archives (Archives) and the National Heritage Board (Museums) of Singapore. Archives' and Museums' metadata records are ingested into NLB's repository through the process of mapping, crosswalking and harvesting. All records can be searched through NLB's OneSearch, which is an integrated discovery service developed by NLB for the searching of physical and digital resources.

2. Data Preparation

Libraries, Archives and Museums metadata records share some common fields but largely use fields which are unique to their collections. NLB uses MARC21 for its physical resources and Dublin Core Libraries Application Profile (DC-Lib) for its digital collections. While Archives uses ISAD-G schema for archival description, Singapore's museums uses its own localised schema. As such, there is a need to harmonise these schemas so that seamless search can occur. Archives and libraries organise materials differently. The multi-level description of archives relates objects in a hierarchy and links the parts to a larger ensemble from the collection level perspective. Libraries organise at the item-level and groups these into collections for discovery. The museums' granularity of description for descriptive areas such as materials & techniques, styles and period, etc. are mostly not found in the descriptions for libraries and archives. Nonetheless, to achieve OneSearch, NLB takes the approach of crosswalking the various schemas to Dublin Core. The crosswalked records from MARC21, ISAD-G and the local schema are ingested into NLB's repository.

3. Controlled Vocabularies and Name Headings Integration and Standardisation

Like other NLs, NLB uses names authorised by Library of Congress Name Authority Cooperative Program (NACO) which observe strict rules for capturing every part of a name. Where names cannot be established in NACO, NLB uses a separate list from a local file. Archives has 8 databases comprising Posters; Oral History Interviews; Government Records; Audiovisual Recordings; Photographs; Maps & Building Plans; Straits Settlements, Overseas & Private Records; and Speeches & Press Releases. These databases are managed by different teams of archivists who do not necessarily share name headings or controlled vocabularies. Whereas Singapore's museums do not use any controlled lists at the time of the project. In the Archives & Museums portal prior to OneSearch, a single person may have more than one form of name. Search results for this person will not be unified and resources are retrieved according to the name a user enters. Merging of Archives collection as it is with the NLB collection will cause

search results to be more fragmented. A search for an entity whether a person, an organisation or a place will not pull content resources about the entity into a single list.

The mapping of vocabularies between the three collections created a consolidated list of controlled vocabularies and name headings shared by NLB, Archives and Museums. The controlled lists of terms and name headings are managed using NLB's vocabulary editor (Taxonomy & Thesaurus Editor). Upon receiving name headings provided by Archives and Museums, NLB's team performed term-matching with NLB's controlled terms and names. Unmatched names are created as new records. For both matched and unmatched names, the team needed to reconcile differences in the form of name used and come to an agreement on an authorized format with variant forms captured in the record. Material types between NLB and Archives were similarly mapped, merged and added to. For the museums, a study was made of the Art and Architecture Thesaurus (AAT) and recommendations were made on terms to be adopted for object categories and material types. These were then merged into NLB's controlled list. The standardized terms reduce ambiguity and increase precision in searching. As a result, metadata of Archives and Museums content, regardless of format, can now be described consistently among NLB, Archives and Museums.

NLB uses ANSI/NISO Z39.19 for the construction, format and management of controlled vocabularies in the Taxonomy & Thesaurus Editor (TTE). To achieve standardization, NLB's TTE will be integrated to the Archives' indexing system. In addition, NLB has shared its policies and guidelines on name creation as well as an agreed upon, workflow between NLB and Archives. Training has been provided to Archives staff. Upon completion of the integration, further training will be provided to ensure processes are standardised so that the three agencies will be able to create and use a shared set of vocabularies and headings.

4. Technical Development

Five and one-half million NLB and 800,000 NAS metadata records were harmonised and OneSearch was launched in August 2014. On 3 June 2015, NLB completed the harmonisation of 80,000 museums metadata and enabled these for searching on OneSearch. The following were some of the technical developments:

1. OneSearch, the integrated search and online interface of NLB, Archives and Museums collections. This includes nine high-level groupings or containers of resources from the 3 agencies namely Books, Magazines & Articles, Audiovisuals, Images, Newspapers, Records & Papers; Websites, eJournals, and Physical Objects.
2. Implementation of search engine optimisation enhancements to improve the discovery of Archives and Museums content using the popular Internet search engines.
3. Deployment of embedded search service in NLB sites and cross-linkages within existing NLB and Archives content sites. Embedment of search service in Museums content sites is work in progress.

5. Benefits

The project was completed in August 2014. However, even before it was officially launched, the page view of Archives had seen substantial jump. During the period between April and November 2014, the page view of Archives was 2.1 million, nearly five times of FY14 target of 450,163. The increase of page views is largely attributed to the re-design of NAS website for search engine optimization and the launch of OneSearch in August 2014.

Users have also directly benefited as they can now access a wider range of materials from NLB, Archives and Museums through a single search without the need to go to three different websites.

The Data Harmonisation has received extensive media coverage as OneSearch is the first search in Singapore which allows users access to a wide range of materials from Singapore's

Libraries, Archives & Museums, including digitised newspapers, literary works, museum artefacts, paintings, manuscripts and speeches, apart from the usual collection of books, magazines and audio-visual materials.

Arquitetura Semântica de Recuperação da Informação

Caio Saraiva Coneglian
UNESP - Universidade
Estadual Paulista, Brasil
caio.coneglian@gmail.com

Elvis Fusco
UNIVEM - Centro
Universitário Eurípides de
Marília, Brasil
fusco@univem.edu.br

José Eduardo Santarém
Segundo
USP - Universidade de São
Paulo, Brasil
santarem@usp.br

Keywords: ontologia; recuperação da informação; web semântica; metadados

1. Introdução

A explosão de geração massiva de dados está testando a capacidade das mais avançadas tecnologias de armazenamento, tratamento, transformação e análise de informações. As áreas do tratamento e da recuperação da informação estão sendo desafiadas pelo volume, variedade e velocidade de uma inundação de dados semi-estruturados e não estruturados de natureza complexa, que também oferece às organizações excelentes oportunidades de ter um aprofundamento no conhecimento mais preciso de seus negócios.

Neste contexto, surgem inúmeras oportunidades em agregar valor ao negócio com base nessas informações, que são geradas tanto no ambiente interno quanto no externo, porém há a necessidade de uma nova abordagem na estrutura de TI (Tecnologia da Informação) das empresas em transformar esses dados em conhecimento para as organizações, que causará impacto de longo alcance.

Desta forma conseguir recuperar dados espalhados na Web de modo que estão possam ser interoperáveis se torna fundamental para conseguir inserir valor a estes dados. Pois ter grandes quantidades de dados desestruturados espalhados pela Web não significa ter dados que tenha valor às organizações.

Para tanto, a Web Semântica (Bernes-Lee et al, 2001) tem dado uma nova direção para como a web deve ser estruturada. De modo geral, a Web Semântica tem como proposta a criação de uma nova estrutura de armazenamento de dados, separando a estrutura de apresentação e do conteúdo das informações (Santarem Segundo, 2014).

Desta forma, os metadados possibilitam que as informações possam ser rotuladas, de modo a, permitir que o computador consiga entender o significado daquela informação, não ficando preso a como aquela informação está sendo apresentada ao usuário. Esta pesquisa, assim, utiliza dos conceitos e tecnologias da Web Semântica e de metadados, para possibilitar que a recuperação das informações ocorra de modo mais semântico e inteligente, observando o contexto, na qual as informações estão inseridas.

Portanto, esta pesquisa tem como objeto propor uma arquitetura que realize a recuperação de dados desestruturados de modo a torna-los interoperáveis com outros sistemas, podendo, assim, trazer informações com alto valor agregado às organizações.

2. Arquitetura

A criação de um agente de software que agregue semanticamente as informações disponíveis na Web de um determinado domínio pode trazer para uma plataforma computacional subsídios para a criação de um ambiente informacional de apoio à decisão que dê uma visão mais ampla dos cenários internos e externos das informações de relevância na gestão organizacional.

Neste contexto, entende-se a extrema relevância de utilizar agentes de extração de dados por meio de robôs de busca semântica com a utilização de tecnologias de padrões de metadados, e meios para a realização de interoperabilidade, sendo imprescindível na recuperação,

armazenamento, processamento e uso dos mais variados tipos de informações gerados nestes ambientes de grande volume de dados em cenários de Inteligência Competitiva.

Para tanto foi proposta uma arquitetura de Recuperação da Informação no contexto de Big Data como pode ser visto na Figura 1.

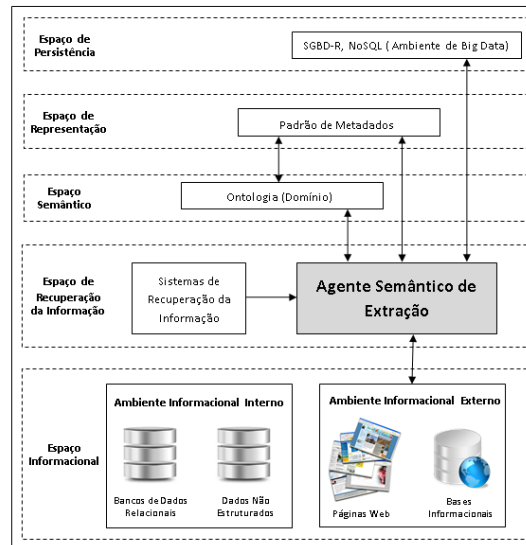


FIG. 1. Arquitetura proposta

No contexto desta arquitetura, esta pesquisa está tratando o problema da extração semântica e automática dos ambientes informacionais na Web que têm como fontes informacionais: páginas Web, serviços Web e base de dados com o desenvolvimento do agente semântico de extração de dados. Este agente deverá se comunicar com os espaços informacionais internos e externos de Big Data baseando suas buscas em regras ontológicas baseadas num padrão de metadados para realizar a extração semântica do domínio proposto e apoiará outros sistemas num contexto mais amplo de Recuperação da Informação.

3. Considerações Finais

Ter acesso às informações do seu domínio de negócio é requisito fundamental para a gestão e à tomada de decisão nas organizações.

Para que um Sistema de Recuperação da Informação tenha a capacidade de disponibilizar as informações relevantes e que estão acessíveis em sites e serviços Web, é necessária a existência de agentes de software que agreguem de forma semântica as informações das mais diversas fontes informacionais de um domínio específico.

Neste contexto os robôs de busca semântica entram como ferramental estratégico na busca e encontro das informações que realmente agregarão valor ao processo decisório, pois dentro de uma imensa e massiva estrutura de dados espalhados pela Web, é imprescindível que os mecanismos de buscas não se apoiem somente em estruturas sintáticas de decisão na recuperação da informação, mas, também em investigações do uso de agentes de extração semântica.

Espera-se que com o uso de uma ontologia de tarefa e um agente semântico de extração agregados em Ambientes de Recuperação da Informação, haja uma efetiva utilização da informação em cenários de Big Data auxiliando ao processo de tomada de decisão.

Referências

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, 284(5), 28-37.
- Santarem Segundo, J. Eduardo. (2014). Web Semântica: Introdução A Recuperação De Dados Usando Sparql. In Encontro Nacional de Pesquisa em Ciência da Informação: além das nuvens, expandindo as fronteiras da Ciência da Informação (pp. 3863-3882,). Belo Horizonte, MG.

Dublin Core: A Metadata Standard in the "3 Marys"

Ana Carla Cunha Nascimento
Universidade de Brasília,
Brasil
anascimento.unb@gmail.com

Rayssa Thaynara Madeira
Correia
Universidade de Brasília,
Brasil
rayssatmadeira@gmail.com

Márcio Bezerra da Silva
Universidade de Brasília,
Brasil
marciobdsilva@unb.br

Keywords: Dublin Core; 3 Marys.

1. Introduction

With recent technological advances, archives, libraries and museums turn to digital environments in order to share, in a wider way, the information contained in their collections. Metadata fit here as resources used for the organization and delivery of content in digital spaces. According to Baptist (2007, p. 181), as a description feature, metadata "help identifying the essential and complementary elements for an effective documentary representation."

The cited discursive context can be met in studies of Information Science (IS), which represents the first step of a survey on the use of metadata standards in digital environments among the "3 Marys" of Smit, lettering representing three areas of IS: archivology, bibliothecology and museology. In this case, from the need for greater understanding of the Dublin Core Metadata Initiative (DCMI) used in the description and mediation of information and their fields, aimed to measure the literary production on DCMI held in "3 Marys" and, specifically, identify the documentary mass by type of material.

2. Materials and methods

Research conducted on the Google search service about DCMI, taking into account up to the third research level and up to 10 results per page. The data collection approach was quantitative between articles to e-zines, blogs, and Web environments for document sharing such as Slideshare, Scribd, Research Gate and Academia.edu, here called clouds. The data been organized in a table, showing the sampling of 30 types of materials identified in each "Marys".

3. Theoretical Foundation

The description and information mediation in digital environments happen through the metadata, which in turn facilitate the import, export and integration of data. The metadata have emerged to help in the organization and retrieval of content available on the web in a growing momentum of rapid and disorderly manner.

It is noticeable the importance of metadata at present, since they allow interaction between systems / digital environments, thus providing not only adequate description and retrieval of information as well as ensures "[...] that resources will survive and continue to be accessible in the future" (NISO, 2004. p. 1). As a metadata example the DCMI stands out, which emerged during the second International Conference on *Web* in Chicago in 1994. This meeting originated the metadata standard to facilitate the description of digital resources through descriptive elements. It is a simple standard with universal semantic understanding and has an extensibility that allows adjustments according to the needs of description. It uses the markup language *eXtensible Markup Language* (XML) and consists of 15 basic elements, namely: title, creator, subject, description, publisher, developer, date, type, format, identifier, source, language, relation,

coverage and copyright. It is noteworthy that the DCMI fields can be implemented (schemes) to the user's discretion and thus enable interoperability with other formats.

4. Results

The production in DCMI was highest in biblioteconomy, corresponding to 50% of the total, followed by archivology (33%) and museology (17%). In this amount, nine, five and three journal articles represented the most identified material type in the "3 Marys" respectively. While results already expected, the DCMI is the most discussed model, widespread and applied in biblioteconomy compared other "Marys". In addition, DCMI is a recurring topic in IS research, especially in current times when technological resources direct the society dynamics and influence the value of information.

TABLE 1: Production at the "3 Marys".

"Marys"	Magazine articles	Others	Cloud	Blogs	Total	Percentage (%)
Biblioteconomy	9	3	2	1	15	50%
Archivology	5	3	2	0	10	33%
Museology	3	1	1	0	5	17%
Total	17	7	5	1	30	100%

5. Final Considerations

The way a "Mary" treats the information in its digital environment can be of great help and for learning to others that have similar features in their digital environments. For this, the XML markup language is the one that has been most widely adopted in these digital environments. It were concluded that the DCMI could be considered an effective finding aid and mediation of information while improving the recovery of information in environments such as digital repositories, especially studied in the library, using the magazine articles as a literary production spaces (scientific).

References

- BAPTISTA, D. M. (2007, Julho/Dezembro). O impacto dos metadados na representação descritiva. *Revista ACB: Biblioteconomia em Santa Catarina*, 12(2). Recuperado, Julho 11, 2015, from <http://revista.acb.org.br/racb/article/view/529>.
- LOURENÇO, C. A. (2007, Janeiro/Fevereiro). Metadados: o grande desafio na organização da Web. *Informação & Sociedade: estudos*, 17(1). Recuperado, Julho 10, 2015, from <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/466/1466>.
- NATIONAL INFORMATION STANDARDS ORGANIZATION. (2004). *Understanding Metadata*. Retrieved, Março 17, from <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>.
- ROCHA, R. P. Metadados – Esquemas. (2012). Recuperado, Julho 10, 2015, from http://www.ufrgs.br/snote/wiki/doc.php?u=ensino/notasaula/metadados_-_esquemas.
- SMIT, J. W. (2003, Junho/Dezembro). *Arquivologia/Biblioteconomia: interfaces das ciências da informação*. *Informação & Informação*, Londrina, 8(1). Recuperado, Julho 11, 2015, from <http://www.uel.br/revistas/uel/index.php/informacao/article/view/1713>.

Particle Physics Metadata Standards in the Tritium File Format

Kevin Wierman
University of North
Carolina at Chapel Hill, &
Metadata Research Center
Drexel University, USA
kjlw@unc.edu

Adrian Ogletree
Metadata Research Center
Drexel University, USA
aogletree@drexel.edu

Jane Greenberg
Metadata Research Center
Drexel University, USA
janeg@drexel.edu

Keywords: metadata; particle physics; Tritium

This poster reports on a standard under development for particle physics data description. This standard is enforced under a file format called Tritium, which is being developed in a framework of the same name.

1. Particle Physics and Sparse Metadata Interfaces

Currently in the field of particle physics, no universal standard exists for metadata between experiments. Individual experiments such as the ATLAS experiment (Malon et al., 2012) and the STAR experiment (Arkhipkin et al., 2015) have developed their own metadata frameworks. The astronomy community has a history of shared metadata practices and standards. (Feigelson et al., 2012). However, in stark contrast, the metadata infrastructure supporting Big Data in particle physics, spanning multiple experiments and collaborations, is limited.

A chief reason for a lack of infrastructure is that the needs of Big Data are contrary to the needs of particle physics experiments. Raw data in particle physics are recorded at high rates due to the volume of data necessary to reconstruct particle interactions. Metadata, therefore, are usually kept sparse in order to allow for faster recording of raw data. In addition, the software that most experiments depend on assumes an existing data structure; and this requires the original software to decode. Big Data requires sufficient metadata (Drake, 2011) to reconstruct experiments and should depend on standards instead of pre-existing software to present data to the community. This goal aligns with developing research on reproducible research (Akmon et al., 2011; Borgman, 2012).

2. Tritium: A multilevel API for file format development

In order to explore the metadata needs of the particle physics community, we are developing a platform designed to address the requirements of both particle physics and Big Data. This platform, Tritium, is comprised of three API's. Each [component?] is designed to address the problems of development in hierarchal manner. Protium, the first level API, is meant to address the data rate and networking requirements for the file format. Deuterium, the second level API, is designed to address the completeness requirements for metadata. Tritium, the top level API, is meant to address the universality requirement for the platform.

Before speaking of the platform itself, it should be noted that this platform is designed to be compatible across operating systems. Figure 1 shows the platform's data explorer working on Android, and Figure 2 shows the platform's metadata explorer working on Linux via python. In this example, the experiment is run on the Android device. The user sets up the experiment using the Tritium interface. The metadata required to describe the data generated from the experiment is recorded by the Deuterium interface and the Protium interface pipes the metadata along with the data to a network output (also described in the metadata) to a Linux machine. The remote machine then adds itself to the metadata description and transcribes the data to file. This illustrates multiple machines and operating systems using the API to record the entire experiment.

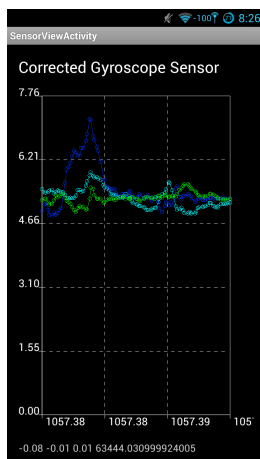


FIG. 1: Tritium Data Explorer in Android (Droidium)

```
<?xml version="1.0" encoding="UTF-8" ?>
<run>
  <start_time time="19700101T000000America/Atikokan(0,0,0,-1,10000)" />
  <stop_time time="19700101T000000America/Atikokan(0,0,0,-1,10000)" />
  <sources>
    <SensorSource maxRange="16384.0" minDelay="5000" name="Corrected Gyroscope Sensor"
ring="(Sensor name="Corrected Gyroscope Sensor"; vendor="Google Inc.")" />
  </sources>
  <encoders>
    <data_encoder name="Corrected Gyroscope Sensor_onSensorChanged" id="28">
      <field name="time" type="Long" />
      <field name="xvalue" type="Int" />
      <field name="yvalue" type="Int" />
      <field name="zvalue" type="Int" />
    </data_encoder>
    <data_encoder name="Corrected Gyroscope Sensor_onAccuracyChanged" id="29">
      <field name="type" type="Int" />
      <field name="version" type="Int" />
      <field name="min_delay" type="Int" />
      <field name="type" type="Int" />
      <field name="max_range" type="Int" />
      <field name="power" type="Int" />
      <field name="resolution" type="Int" />
    </data_encoder>
  </encoders>
</SensorSource>
</sources>
<outputs/>
</run>
```

FIG. 2: Tritium Metadata Explorer in Python

3. Particle Physics Data and MetaData

Particle physics data is comprised of a series of *records*. An individual record may contain multiple fields such as time, sensor value, and sensor number. In addition, particle physics electronics are designed to give addresses to sensors so that data may be traced back through groups of sensors to an individual sensor. The fastest way to transcribe these values is to simply take the raw binary and write it to file. However, data sets may contain mixed record types. The requirement for decoding mixed record types is record type identification. Therefore, it is common practice to assign a dictionary of integers to decoders and transcribe the associated decoder value with each record. In this way, data may be decoded post-data taking in a reproducible way. Even with this approach, there is a frequent shortcoming, as the dictionary is often assumed by the software and not included with the data; this results in an incomplete data set.

Tritium takes this approach a step further by recording the dictionary of the decoder to integer values in a file header. The dictionary includes a description of each field by its data type (integer, float, double, etc.), and a name for the field in order to be identifiable by the user, as in Figure 2. The output file is readable by any number of software platforms and the user only requires the file itself and knowledge of the standard used to create it. While this approach may not be entirely complete, it represents a critical step taken in the right direction.

4. Status and Next Steps

Currently, the Tritium file format includes standards for taking and decoding data. However, for a completely reproducible experiment, the next step in developing Tritium would be to include ways to describe common analysis techniques to file so that the life-cycle of the data can be made available to users at all times.

The other next step is to deploy this platform in other programming languages and operating systems. As displayed, Tritium is operational on Android and Linux, as well as Linux-like systems such as Raspberry Pi's, and lower systems such as Arduinos. Future work will be to make this truly platform-independent.

References

- Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. *Archival Science*, 11(3-4), 329-348.
- Arkhipkin, D. & Lauret, J. (2015). Journal of physics. conference series: STAR Online Framework: from Metadata Collection to Event Analysis and System Control Institute of Physics. doi:10.1088/1742-6596/608/1/012036

- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059-1078. doi:10.1002/asi.22634
- Drake, C. (2011, Nov 24). Turning regulatory big data requirements on their head. *City A.M.* Retrieved from <http://search.proquest.com/docview/905854534?accountid=14244>
- Feigelson, E. D., Feigelson, E. D., & Babu, G. J. (08/01/2012). *Significance* (oxford, england): Big data in astronomy Blackwell Publishing. doi:10.1111/j.1740-9713.2012.00587.x
- Malon, D., Albrand, S., Gallas, E., & Stewart, G. (2012). *Journal of physics. conference series: A programmatic view of metadata, metadata services, and metadata flow in ATLAS Institute of Physics.* doi:10.1088/1742-6596/396/5/052052
- Shapoval, I., Clemencic, M., & Cattaneo, M. (2014). *Journal of physics. conference series: ARIADNE: A tracking system for relationships in LHCb metadata Institute of Physics.* doi:10.1088/1742-6596/513/4/042039

Use and Connect: Linked Open Data of the National Diet Library, Japan

Yoshikazu Nagai
National Diet Library,
Japan
ysnagai@ndl.go.jp

Akiko Hashizume
National Diet Library,
Japan
hasizume@ndl.go.jp

Julie Fukuyama
National Diet Library,
Japan
ju-fukuy@ndl.go.jp

Keywords: The National Diet Library; Linked Open Data; Japanese library, metadata

1. About NDL

The National Diet Library (NDL) is the sole national library in Japan. The NDL acquires, preserves and provides Japanese publications which are the nation's cultural and intellectual assets. The acquisition of library materials is mostly based on the Legal Deposit System. The NDL compiles and provides various bibliographies of library materials. Most of the collections are searchable through the NDL-OPAC and NDL Search on the website. To facilitate effective data use by computer systems or applications, the NDL initiatives to promote Linked Open Data (LOD) and provides metadata as LOD.

2. What is NDL LOD?

The NDL provides LOD of bibliographic data (NDL Search), authority data (Web NDL Authorities), earthquake related data (NDL Great East Japan Earthquake Archive (code name "HINAGIKU") and beta version of International Standard Identifier for Libraries and Related Organizations (ISIL) LOD.

2.1. Bibliographic Data (NDL Search)

NDL Search is an integrated information search service that serves as a gateway to the rich repository of knowledge contained in the NDL, public libraries, academic libraries, archives, museums, and academic research institutions in Japan. It officially opened to the public on January 2012 and can search about 83 million metadata records as of March 2015. Data sources for the NDL Search include: NDL-OPAC, Japanese Periodicals Index, National Diet Library Digital Collections, digital archives provided by public and academic libraries in Japan, etc.

NDL Search provides bibliographic data in RDF/XML, and these bibliographic data are of books, journals, articles, newspapers, digital contents (digitized materials, sounds, web pages etc.). These data include title, author, publisher, subject matter, classifications, ISBN, ISSN, National Bibliography No., NDLJP which is used to identify digitized content of the NDL digital collection, URLs of webpages which show digitized content ([http://dl.ndl.go.jp/...](http://dl.ndl.go.jp/)), information related to copyright protection and so on. The National Diet Library Dublin Core Metadata Description (DC-NDL) is used for metadata description.

The NDL Search provides an API (SRU, SRW, OpenSearch, Z39.50 and OAI-PMH) to download data.

2.2. Authority data (Web NDL Authorities)

Web NDL Authorities is the name of a service that provides NDL authority data as LOD. The service officially started in January 2012. It provides access to about 1.17 million pieces of data as of March 2015. The authority data is information which identifies authors who have several names (pen names, maiden names, etc.) as well as synonyms, information which identifies different people with same names, information on synonyms which indicate a certain topic (subject matter),

hypernyms, hyponyms, related words and so on. The NDL uses SKOS-XL, SKOS, DC-NDL, RDFS, Dublin Core, FOAF and OWL as terms for authority data descriptions.

Web NDL Authorities provides an API via SPARQL to download data. In addition, there are two formats in files for batch download: RDF/XML and TSV. These files contain data of the National Diet Library List of Subject Headings (NDLSH).

2.3. Earthquake related data (NDL Great East Japan Earthquake Archive (code name “HINAGIKU”))

HINAGIKU is a portal site that enables integrated searches of multiple resources on earthquakes and subsequent disasters. The website officially opened to the public in March 2013. A user can search about 2.88 million metadata records as of March 2015. Data sources of HINAGIKU are various institutional repositories which compile records about the Great East Japan Earthquake such as Geospatial Information Authority of Japan, Center for Remembering 3.11 (sendai mediateque) and so on. HINAGIKU provides the following metadata as LOD:

- metadata of photos (aerial photographs of the disaster area and photos of damage), sound recordings/videos (related to support activities for the affected areas, the Fukushima Daiichi nuclear disaster, and testimony of disaster victims, etc.)
- metadata of old web pages (municipal governments, etc.)
- metadata of books, journals, newspapers, and brochures
- URLs of webpages that show digitized content (documents, photos, sound recordings/movies), thumbnail URLs, and URLs of webpages which shows past webpages and so on.

HINAGIKU provides API via SRU, OpenSearch, and OAI-PMH to download data.

2.4. Beta version of ISIL LOD

International Standard Identifier for Libraries and Related Organizations (ISIL) are identifiers which can be allocated to libraries and other relevant organizations, such as archives and museums. ISO 15511 specifies that the ISIL structure be administrated by national allocation agencies in each country. The NDL operates the National Agency for ISIL in Japan.

Since April 2015, we have been operating a beta version of the ISIL LOD. This LOD includes about 7,500 ISIL data, comprising ID, institution name, address, and other items. This dataset also includes longitudes and latitudes of addresses, name authority URIs of the Web NDL Authorities etc. In addition, we have defined the LibType (dcndl:LibType) vocabulary for describing the types of libraries, and it is used within the ISIL LOD.

This dataset is provided as a downloadable file in RDF/XML. There are no restrictions on use of these public domain datasets.

3. Use cases

NDL LOD is used via various systems or applications. The following are typical examples of how NDL LOD are used.

3.1. calil.jp

Managed by CALIL Inc., calil.jp is an online service that enables cross-searching of Japanese library OPACs. By specifying geographical regions before searching, patrons can verify whether or not the books they wish to borrow are available at a library in the specified area. Although calil.jp initially used Amazon metadata for linking with data of library holdings, this method did not allow patrons to find books that are not available at Amazon. To solve this problem, calil.jp now makes significant use of bibliographic data from NDL Search API.

3.2. VIAF

Virtual International Authority File (VIAF) is a system that is managed by the Online Computer Library Center, Inc. (OCLC) and which links multiple name authority data from national libraries and organizations to create cluster data for each unique name. The NDL has participated in the VIAF since October 2012, and the VIAF includes the URI from NDL authority data, e.g. <http://id.ndl.go.jp/auth/entity/00104237>. Web NDL Authorities also provides links to the VIAF. Hence, authority data from the NDL or other organizations are linked through the VIAF, which enables users to find Japanese authority data even when searching in languages other than Japanese.

4. Towards further development

We think there are three issues to be solved in promoting the NDL LOD.

The first is to find solutions to the difficulties inherent in utilizing LOD. Japanese engineers have pointed out obstacles, such as the ambiguity in the terms of use of NDL data as well as the fact that there is no sample code available for the NDL API. In particular, we think it is important that our data be available through an open license.

The second is to enhance the quality of linked data. This involves several underlying issues, such as the fact that some HTTP URIs in our LOD cannot refer to these values, and our LOD includes few links to outside datasets.

The third is to provide other kinds of data as LOD. As the sole national library in Japan, the NDL is expected to provide metadata vocabularies, which are necessary to convert Japanese bibliographic data into LOD. Our first step in this process involved publication of a beta version for ISIL LOD. At this time, we are trying to convert the Nippon Decimal Classification (NDC), the standard classification system in Japan, to linked data.

References

- Beta version of International Standard Identifier for Libraries and Related Organizations (ISIL) LOD. <http://ndl.go.jp/en/aboutus/standards/opensdataset.html>
- Calil. Retrieved June 08, 2015, from <http://calil.jp/>
- DC-NDL. Retrieved June 08, 2015, <http://www.ndl.go.jp/en/aboutus/standards/meta.html>
- NDL Great East Japan Earthquake Archive (HINAGIKU). Retrieved June 08, 2015, from <http://kn.ndl.go.jp/node?language=en>
- NDL Search. Retrieved June 08, 2015, from <http://iss.ndl.go.jp/?locale=en>
- VIAF. Retrieved June 08, 2015, from <http://viaf.org/>
- Web NDL Authorities. Retrieved June 08, 2015, from <http://id.ndl.go.jp/auth/ndla>



São Paulo, Brazil