Metadata Intersections:
Bridging the Archipelago of Cultural Memory

**2014 Proceedings of the International
Conference on Dublin Core and Metadata
Applications**

Proceedings Edited by:
William Moen
College of Information, University of North Texas, United States
Amy Rushing
University Libraries, University of Texas at San Antonio, United States

Published by:
Dublin Core Metadata Initiative
a project of ASIS&T

Conference Host:

TDL.ORG
TEXAS DIGITAL LIBRARY

**Austin, Texas, USA
8-11 October 2014**

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

ISSN: 1939-1366 (Online)

DC-2014 Welcome

Welcome to DC-2014 in Austin, Texas! This gathering of researchers, practitioners and students of metadata for the annual meeting and conference of the Dublin Core Metadata Initiative (DCMI) marks the twenty-second formal meeting of our community. It also marks the end of a year of reinvention and reimagining of the way DCMI works, manifested in large part through a return to the organizational model found in the early years of the initiative, with responsibility for direction and management resting in the hands and minds of the membership.

Much of the groundwork for this re-envisioning came from meetings of the Advisory Board held at the annual meeting in Lisbon last year, and has been shepherded through early stages by the newly elected officers of the Advisory Board, the newly named Governing Board (formerly the Oversight Committee), and the newly formed Technical Board, along with critical assistance from our current Managing Director, Stuart Sutton. We believe the end outcome will be a stronger, member-driven organization that opens new doors for ideas and initiatives that will build on the strengths and reputation DCMI has already established in the international community.

The program this year has a number of new innovations designed to foster this goal, ranging from the Best Practice Posters and Demonstrations, which will showcase concrete examples of current practice in metadata applications, to the Next Generation Metadata Specialist Program, which provides an opportunity for emerging professionals to network with veterans throughout the conference. As always, there are many opportunities to catch up with some of the boundary-pushing technical work being accomplished by task groups and your colleagues in their everyday work, as well as pre and post conference workshops and tutorials. Many thanks to the program committee and chairs for creating a stimulating and diverse program this year.

As you participate in this year’s conference we hope that you will think about how you can contribute to the growth and strengthening of DCMI in the coming year—through contributions to the technical work and outreach being accomplished by task groups, by volunteering to take on a role as a chair or co-chair of one of the Standing Committees or the Technical or Advisory Board, by helping engage new members or re-engage old participants, or just by joining as an Individual Member to help support the important work DCMI is doing that we all benefit from in our own activities during the rest of the year.

I personally hope that many of you will also take advantage of the opportunity to add your voice and support to the background work critical to keeping DCMI alive and thriving by attending the Annual Meeting on Saturday. Active members are important in keeping the initiative moving ahead, and this is a chance to join in that work and meet some of the people who contribute their time and effort to making it possible for all of us to reap the benefits of the thought and creativity engendered through DCMI’s activities.

We are sure that you will find the conference and meeting exciting and that you will leave Austin with an even greater commitment to the work DCMI is doing, and a deeper engagement with your colleagues in the world of metadata throughout the coming year. Enjoy!!

Michael Crandall, Chair, DCMI Governing Board

Chair's Notes on the Program

At the 2012 South by Southwest Interactive Conference in Austin, Texas, Jon Voss led a panel of speakers to introduce and discuss a "global movement afoot" that encourages greater public access to metadata in the world's libraries, archives, and museums. The movement, led by a network of practitioners and professionals across cultural heritage institutions, aims to increase adoption and implementation of Linked Open Data within the cultural heritage community. The panelists discussed use cases and applications of linked open data and presented a variety of possibilities for cultural data access, remix, and reuse.

In keeping with the Dublin Core's history of reflecting and engaging the evolution of the metadata field, this year's conference builds upon that movement Voss and his panelists spoke about. The theme of this year's conference – "Metadata Intersections: Bridging the Archipelago of Cultural Memory"—acknowledges that while metadata is the essential element to enabling access to the world's galleries, libraries, archives, and museums (GLAM), there are significant differences in domain praxis. The conference program explores how these differences may be bridged in the context of linked open data.

A pre-conference comprising a full-day workshop and two half-day tutorials launches the conference. A post-conference workshop on linked data brings the week to an end. These enable deeper engagement with a variety of topics that touch on this year's theme including emerging practices in archival description; linked open data hands on training; RDF in the cultural heritage sector; and a historical overview of the accomplishments of the DCMI community.

This year's conference program includes two days of full-length conference papers and project reports. Special sessions and poster viewings run concurrent to the papers and reports throughout the two days. Participants from a variety of cultural heritage institutions and practitioners utilizing linked open data and semantic web technologies will present both theoretical and project-based papers. In this year's submissions, we are seeing true momentum in the exploration and adoption of linked open data across all cultural heritage sectors.

DC 2014 unveils two new efforts: one that attempt to recruit young professionals and students to attend, and the other to provide more opportunities for presenting the best practices of metadata workers. The Next Generation Metadata Specialist Program solicited iSchools, other library and information science programs, and libraries to sponsor one or more of their students and early-career metadata professionals to attend the conference. Thirteen organizations are participating. The participants selected for the Next Generation Metadata Specialist Program will engage one on one and in group interactions with leading researchers, consultants, and practitioners shaping the metadata ecosystem and in a special session, designed for them; they will gain an understanding of how the discourse and practice of metadata are evolving.

Also new this year, the non-peer reviewed Best Practices Poster and Demonstrations tracks. Intended to encourage practitioners to showcase innovative approaches to metadata best practices, these tracks garnered a great response: we have a total of 17 posters and two demonstrations. A conference as special as Dublin Core owes so much to so many. We are grateful to all of the people who submitted proposals to share their ideas, experiences, and research. Similarly we are grateful to the many people who volunteered their time as reviewers of all of those proposals. As program co-chairs we are especially grateful for the opportunity to serve and contribute to this year's conference.

William E. Moen, College of Information, University of North Texas, United States
Amy Rushing, University of Texas at San Antonio Libraries, United States

ORGANIZING COMMITTEE

DC-2014 Conference Committee Chair

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

Program Committee Chairs

William E. Moen, College of Information, University of North Texas, United States
Amy Rushing, University of Texas at San Antonio Libraries, University of Texas at San Antonio, United States

Outreach Committee Chair

Eric Childress, OCLC Research, United States

Local Organizing Committee Chair

Kristi Park, Texas Digital Library

Program Committee

Leif Andresen, Advisor to the Director the Royal Library. National Library of Denmark, Denmark
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
Marie-Claude Côté, Treasury Board Secretariat of Canada, Canada
Karen Coyle, Consultant, United States
Michael D. Crandall, University of Washington
Makx Dekkers, AMI Consult SARL, Spain
Jacques Ducloy, University of Lorraine, France
Gordon Dunsire, Independent Consultant, United Kingdom
Kai Eckert, University of Mannheim, Germany
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, University of North Carolina, Chapel Hill, United States
Willem Robert van Hage, VU University Amsterdam, Netherlands
Corey A. Harper, New York University
Seth van Hooland, Université Libre de Bruxelles, Belgium
Eero Hyvönen, Aalto University, Finland
Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Netherlands
Masahide Kanzaki, Keio University Xenon Limited Partners, Japan
Tomi Kauppinen, University of Muenster, Germany
Johannes Keizer, Food and Agriculture Organization of the United Nations (FAO), Italy
Dean Blackmar Krafft, Cornell University Library, United States
Michael Lauruhn, Elsevier, United States
Akira Maeda, Ritsumeikan University, Japan
Filiberto Felipe Martinez-Arellano, National Autonomus University of Mexico, Mexico
Philipp Mayr, GESIS - Leibniz Institute for the Social Sciences, Germany
Eva M. Méndez, University Carlos III of Madrid, Spain
Shawne Miksa, University of North Texas
Steven J. Miller, University of Wisconsin-Milwaukee School of Information Studies, United States
Akira Miyazawa, National Institute of Informatics, Japan
William E. Moen, College of Information, University of North Texas, United States
Peter E Murray, LYRASIS, United States
Jin-Cheon Na, Nanyang Technological University, Singapore
Liddy Nevile, Independent Consultant, Australia
Annelies van Nispen, Eye Film Institute, Netherlands

Johan Oomen, Netherlands Institute for Sound and Vision, Netherlands
Jung-ran Park, College of Information Science and Technology, Drexel University, United States
Oknam Park, Sangmyung University, Republic of Korea
Cristina Pattuelli, Pratt Institute, United States
Vivien Petras, Humboldt-Universität zu Berlin, Germany
Magnus Pfeffer, Stuttgart Media University, Germany
Serhiy Polyakov, University of North Texas, United States
Sarah Potvin, Texas A&M University Libraries, United States
Jian Qin, Syracuse University, United States
KS Raghavan, Centre for Knowledge Analytics & Ontological Engineering, PES Institute of Technology, India
Stefanie Ruehle, SUB Goettingen, Germany
Amy Rushing, University of Texas at San Antonio, United States
Johann Wanja Schaible, GESIS - Leibniz-Institute for the Social Sciences, Germany
Bernhard Schandl, Gnowsis.com, Austria
Jodi Schneider, INRIA Sophia Antipolis, 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
Joseph T. Tennis, University of Washington, United States
Douglas Tudhope, University of Glamorgan, United Kingdom
Vassilis Tzouvaras, National Technical University of Athens, Greece
Sherry L. Vellucci, University of New Hampshire, United States
Paul Walk, EDINA, United Kingdom
Mei-Ling Wang, Graduate Institute of Library, Information and Archival Studies, Taiwan
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

Local Organizing Committee

Linda Abbey, University of Texas Libraries
Effie Bradley, Texas Digital Library
Debra Hanken Kurtz, Texas Digital Library
Gad Krumholz, Texas Digital Library
Nick Lauland, Texas Digital Library
Jason Sick, University of Texas Libraries
Ryan Steans, Texas Digital Library
Antoinette Yost, Texas Digital Library

TABLE OF CONTENTS

Distributed metadata Environments & Aggregation—Part A

Distributed metadata Environments & Aggregation—Part B

Metadata in Support of Research

Infrastructure & Models—Part A

Infrastructure & Models—Part B

Metadata Praxis

Posters (Peer Reviewed)

Best Practice Posters & Demonstrations

Distributed Metadata Environments & Aggregation—Part A

Linked Data Mapping Cultures: An Evaluation of Metadata Usage and Distribution in a Linked Data Environment

Konstantin Baierer
Humboldt-Universität zu Berlin, Germany
konstantin.baierer@ibi.hu-berlin.de

Evelyn Dröge
Humboldt-Universität zu Berlin, Germany
evelyn.droege@ibi.hu-berlin.de

Vivien Petras
Humboldt-Universität zu Berlin, Germany
vivien.petras@ibi.hu-berlin.de

Violeta Trkulja
Humboldt-Universität zu Berlin, Germany
violeta.trkulja@ibi.hu-berlin.de

Abstract

In this paper, we present an analysis of metadata mappings from different providers to a Linked Data format and model in the domain of digitized manuscripts. The DM2E model is based on Linked Open Data principles and was developed for the purpose of integrating metadata records to Europeana. The paper describes the differences between individual data providers and their respective metadata mapping cultures. Explanations on how the providers map the metadata from different institutions, different domains and different metadata formats are provided and supported by visualizations. The analysis of the mappings serves to evaluate the DM2E model and provides strategic insight for improving both mapping processes and the model itself.
Keywords: mapping evaluation; ontology evaluation; mapping varieties; DM2E model; Linked Data; Europeana

1. Introduction

Do mapping preferences of individual institutions influence the resulting data from a mapping process? In this paper, mapped datasets from eight different data providers (DP) processed by six different mapping institutions (MI) were analyzed. The primary aim of the analysis was an evaluation of the model to which the data is mapped. Based on the differences of mappings in the evaluation, different Linked Data mapping cultures emerged.

The evaluation of a dataset or data model provides insight into over- and underused parts of the model or misrepresented or misunderstood data mappings. Previous studies have looked at the distribution and usage of fields or model classes and properties and the mapping data in library catalogs (e.g. Seiffert, 2001; Smith-Yoshimura, Argus et al., 2010). These studies show that only a subset of the provided properties in data formats are used in practice. Palavitsinis, Manouselis & Sanchez-Alonso (2014) observed in their study of metadata quality in cultural collections that the "perceived usefulness for all elements of an application profile drops when the number of these elements rises" (p. 9). In Linked Data research, the focus has been on the analysis of certain vocabularies (e.g. Alexander, Cyganiak et al., 2009) and statistics on individual or aggregations of RDF datasets including data accessibility and coverage (Auer, Demter et al., 2012). Klimek, Helmich & Nacasky (2014) built a Linked Data Visualization Model (LDVM) which creates an analytical RDF abstraction and a visual mapping transformation.

This paper first introduces the DM2E model and its application context and then provides general statistics on the use of different model classes and properties by different providers and mapping institutions. Different data and model characteristics are discussed to provide an analysis of different mapping styles (cultures) and their consequences.

2. A Data Model for Cultural Heritage

Europeana[1] is the European digital library, which gives access to more than 30 million library, archive, museum and audio-visual objects from 36 countries. These objects are digitized and described by content providers in different metadata formats. National or domain aggregators deliver the object metadata to Europeana in the Europeana data model (EDM) (EDM Primer, 2013). Digitised Manuscripts to Europeana (DM2E)[2] is a domain aggregator contributing to the development of Europeana. Among other goals, DM2E collects, maps and delivers rich metadata about manuscripts to Europeana.

The metadata mapping and the ingestion of mapped data into Europeana are supported by a specialization of the EDM for manuscripts that was developed for DM2E. The EDM is very broad and generic in order to fit the different metadata standards like TEI or METS/MODS in which cultural heritage objects (also referred to as CHOs) are described by data providers. The model is RDF-based and can thus easily be extended by others as done in the DM2E project. The resulting specialization is called the DM2E model.

The DM2E model (Dröge, Iwanowa & Hennicke, 2014a) has been built as a specialization of the EDM in order to represent rich manuscript metadata in Europeana, which is also published as Linked Open Data (LOD) (Heath & Bizer, 2011). The development approach of the model was bottom-up: requirements from data providers as well as from technical partners were collected and new properties or classes were created or reused from external vocabularies. Properties and classes were added as subproperties / -classes to EDM resources when possible in order to enable backwards compatibility. In that way, the main structure of the EDM remains unchanged in the DM2E model. The core classes of both models are *edm:ProvidedCHO* for the cultural heritage object, *ore:Aggregation* for the provided metadata record and *edm:WebResource* for Web resources related to a CHO, e.g. an image of it. The class that is most extensively specialized in the DM2E model is *edm:ProvidedCHO*. More than 50 properties were added to this class to better describe the creator of a CHO, its contributors and concepts, places and time spans related to it. Similar to the EDM, the DM2E model mainly focuses on properties and not on classes to describe the provided data. Nevertheless, a small amount of classes were also added, e.g. to differentiate various types of CHOs like *dm2e:Page*, *bibo:Book* or *fabio:Article*. These classes are important to model hierarchical objects which are not yet fully supported in EDM.

3. Distribution of Classes and Properties

Ten datasets mapped to the RDF-based DM2E model describing manuscripts, books, letters and journal articles were analyzed. The total amount of RDF statements in the analyzed sample is 61,365,146. The data was delivered by eight data providers (DP) and mapped by six different mapping institutions (MI). The DPs, MIs and datasets were anonymized as the focus of the study does not lay in specifics of a single dataset but in the differences between the mapping behaviour of the six MIs. Our assumption is that not only the provided data but also the particular mapping approach influences the resulting data in the DM2E model. Table 1 shows the providers, datasets, the metadata format of the data before the ingestion and the responsible mapping institution. All data was mapped to the DM2E model version 1.1, latest revision (Dröge, Iwanowa et al., 2014b).

[1] Europeana website: http://europeana.eu/ (last accessed 22.04.2014).

[2] DM2E website: http://dm2e.eu/ (last accessed 22.04.2014).

[3] https://github.com/DM2E/dm2e-analysis/tree/master/sparql (last accessed 15.05.2014).

[4] https://github.com/DM2E/dm2e-analysis/blob/master/build_tables.py (last accessed 15.05.2014).

[5] https://developers.google.com/chart/ (last accessed 15.05.2014).

[2] DM2E website: http://dm2e.eu/ (last accessed 22.04.2014).

The first aim of the analysis was to evaluate the DM2E model by identifying properties and classes that were not mapped. Unmapped resources could potentially be removed from the model to reduce its complexity. The analysis of the mappings could also be used to evaluate whether the model can cover different domains. Can a generic model like the EDM and its specializations be used to represent this data or do the Linked Data mapping cultures vary too much? Does a mapping reflect the institution that has mapped the data?

TABLE 1: Analyzed datasets.

Data Provider (DP)	Dataset	Metadata format	Mapping institution (MI)
DP I	Dataset 1	proprietary format	MI A
DP I	Dataset 2	proprietary format	MI A
DP II	Dataset 3	MAB2	MI B
DP II	Dataset 4	MAB2	MI B
DP III	Dataset 5	METS/MODS	MI C
DP IV	Dataset 6	METS/MODS	MI C
DP V	Dataset 7	TEI P5	MI D
DP VI	Dataset 8	EAD	MI D
DP VII	Dataset 9	TEI P5	MI E
DP VIII	Dataset 10	TEI P5	MI F

The evaluation reported in this paper is based on an automated analysis and visualizations. The RDF data in the triple store is organized in Named Graphs (Carroll et al., 2005), each Named Graph representing a specific ingestion of a specific dataset including full provenance. Using SPARQL, the latest ingestion of each dataset was determined. Then, a set of SPARQL queries was run on the data in these ingestions[3] to gather the raw counts for various quantifiable aspects of these datasets, including generic statistics such as number of statements, number of specific predicates, number of different ontologies, ranges of predicates, RDF types, as well as DM2E-specific statistics such as frequency of certain subclasses of *edm:PhysicalThing* or occurrences of predefined statement patterns. A Python script[4] then collated the raw tabular data, calculated means, sums and ratios within and across datasets and produced HTML with embedded SVG using the Google Chart data visualization API[5]. Unprocessed visualizations[6] and the source code[7] are available.

The providers or mapping institutions used a large variety of classes and properties of the DM2E model and produced rich mappings. Still, more than a half of all classes (24 out of 43) and about a third of all properties (47 out of 125) that the model offers were not used by any of the providers. The counts do not include classes and properties that are used for means beyond manuscript metadata, e.g. for external annotation tools or for tracking provenance within the DM2E interoperability infrastructure.

Figure 1 shows the distribution of all properties. The most frequently used properties are *dc:contributor*, *edm:rights*, *dc:format* und *dc:description*. Properties which must be used exactly once occur for each of the ca. 2.1 million CHOs: *dm2e:hasAnnotatableObject* (strongly recommended), *dc:language* (mandatory), *edm:dataProvider* (mandatory), *dc:type* (mandatory), *edm:aggregatedCHO* (connection between the CHO and the aggregation; this is mandatory and must occur once per object), *edm:type* (mandatory), *dm2e:displayLevel* (mandatory). The property *dc:title* is not mandatory and is used “only” 1,722,542 times in 2,134,934 CHOs. The strongly recommended properties were used almost as often as the mandatory ones. A major part

[3] https://github.com/DM2E/dm2e-analysis/tree/master/sparql (last accessed 15.05.2014).

[4] https://github.com/DM2E/dm2e-analysis/blob/master/build_tables.py (last accessed 15.05.2014).

[5] https://developers.google.com/chart/ (last accessed 15.05.2014).

[6] http://data.dm2e.eu/visualize/index.html (last accessed 24.07.2014).

[7] https://github.com/DM2E/dm2e-analysis (last accessed 15.05.2014).

of the properties is used infrequently compared to the number of CHOs, a logical consequence because specific properties just fit particular datasets. About one third of the properties was not mapped. Both, DM2E-specific properties but also EDM properties, were not mapped. Properties from contextual classes, e.g. coordinates of places (*wgs84_pos:lat*, *wgs84_pos:long*), the date an institution started (*rdaGr2:dateOfEstablishment*) or ended (*rdaGr2:dateOfTermination*) are possibly simply missing in the data. SKOS properties like *skos:broader*, *skos:narrower* or *skos:notation* were not mapped. Uncommon properties like *dm2e:levelOfGenesis*, *dm2e:influencedBy* or *dm2e:misattributed* were not mapped even though they were explicitly requested by data providers. The distribution of properties mirrors previous findings from Seiffert (2001), who analyzed MAB fields of title data in libraries and showed that 58.46% of MAB fields for bibliographic data were unused. The same results could be found in an internal statistical analysis of EDM data at Europeana conducted in January 2014, which concluded that 40% of the fields remained unused.

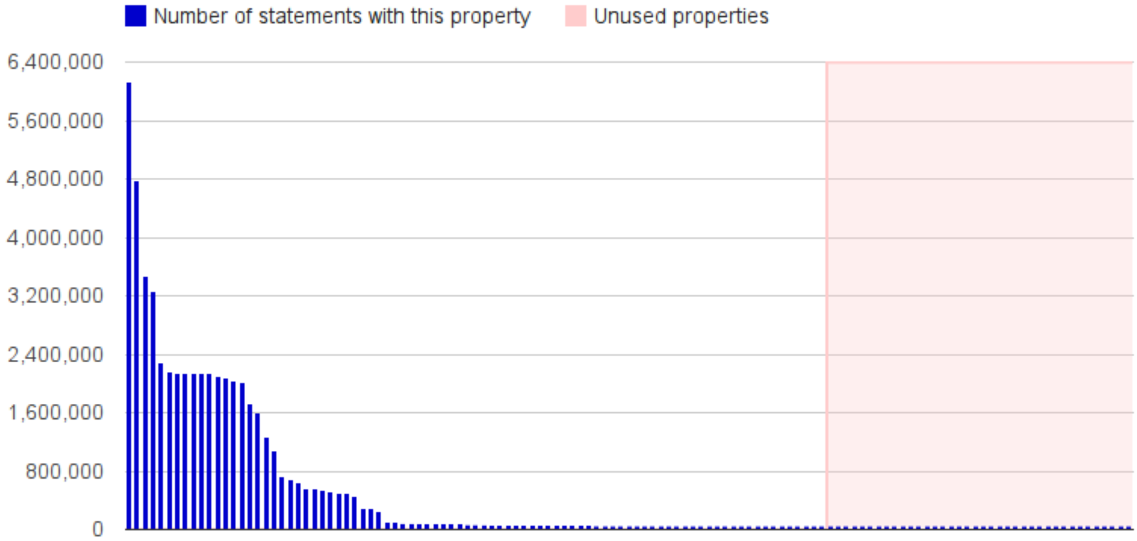


FIG. 1: Absolute frequency of all predicates. Properties on the right side of the vertical bar were never used in any dataset.

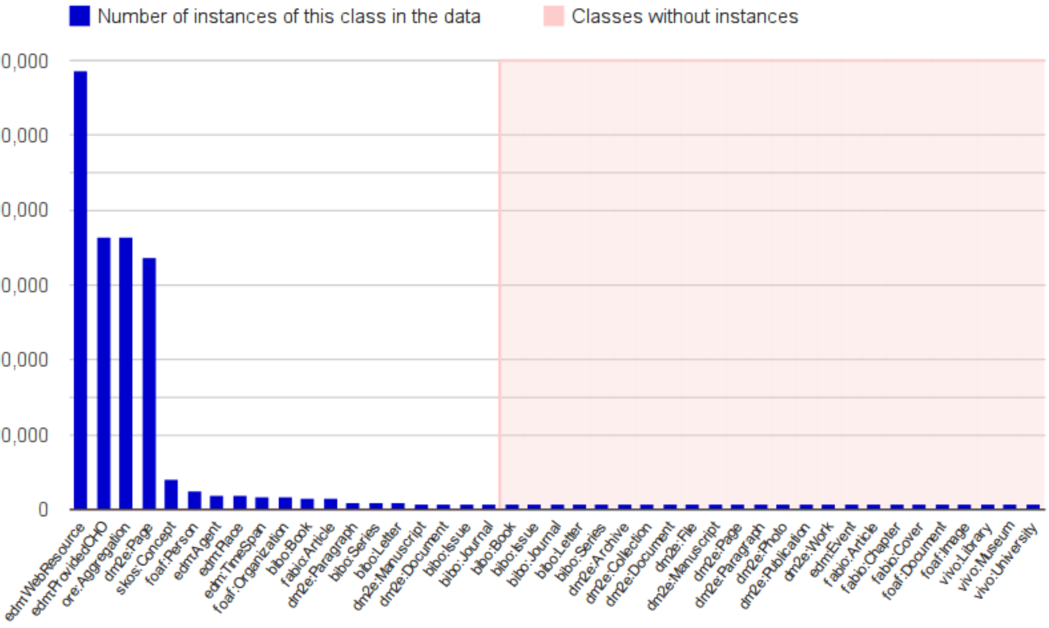


FIG. 2: Distribution of classes across datasets in DM2E.

The most frequently used classes (as shown in fig. 2) are *edm:WebResource* (every CHO must point to at least one Web resource), followed by *ore:Aggregation* and *edm:ProvidedCHO*. They occur equally often, as there is always one aggregation per CHO, and are mandatory. Although contextual classes are not mandatory and less frequently mapped, they are very useful as they allow contextual data to become Linked Data representations with dereferenceable IRIs[8] as opposed to mere strings. The class *skos:Concept* (the fifth most mapped class) is used very unevenly: DP V-Dataset 7 uses it 138,440 times, DP I-Dataset 1, DP III-Dataset 5 and DP II-Dataset 4 do not use it at all. Subclasses of *foaf:Organization*, e.g. *vivo:Library*, *dm2e:Archive*, *edm:Event* were never used. Altogether, 24 of 43 classes are unused.

The class *dm2e:Page* is used most often as the aggregation level of an object (see table 2). While DM2E prepared for different types and aggregation levels, the data appears to be aggregated almost exclusively on the page level. However, in the mappings, several levels are used. Most datasets make use of two different levels of hierarchy within a CHO. This can not only be explained with the provided metadata. For example, chapters are never mapped but exist in the provided books. Which and how many levels of a hierarchical object are mapped seems to be mostly based on the mandatory elements in the model and on the decisions of the MI.

TABLE 2: Different CHO types (subclasses of *edm:PhysicalThing* or *skos:Concept*).

Dataset	**bibo: Series**	**bibo: Book**	**dm2e: Manu-script**	**dm2e: Para-graph**	**bibo: Journal**	**bibo: Issue**	**fabio: Article**	**bibo: Letter**	**dm2e: Page**
Dataset 1	-	-	24	-	-	-	-	-	10,427
Dataset 2		1,251	10						530,314
Dataset 3	4,552	39,873	-	-	-	-	-	-	-
Dataset 4	-	-	175	-	-	-	-	-	46,006
Dataset 5	-	-	1,012	-	-	-	-	-	307,202
Dataset 6	-	2,916	-	-	-	-	-	-	472,994
Dataset 7	-	1,295	-	-	-	-	-	-	416,172
Dataset 8	-	-	-	-	-	-	-	3,630	34,596
Dataset 9	-	-	-	-	1	346	42,173	-	159,277
Dataset 10	-	-	20	9,635	-	-	-	-	-
Total	4,552	45,335	1,241	9,635	1	346	42,173	3,630	1,976,988

Only few mappers use *edm:Agent* (DP IV-Dataset 6: 2,919; DP II-Dataset 3: 11,796; DP VIII-Dataset 10: 35). In the same datasets where *edm:Agent* is used, *foaf:Organization* and *foaf:Person* are mapped as well. *foaf:Organization* and *foaf:Person* are mapped by everyone. In some datasets, they are rarely mapped (DP I-Dataset 1: 2 organizations, 3 persons and 0 agents; DP II-Dataset 4: 0 agents, 33 organizations, 275 persons), in other datasets they are very often mapped (DP II-Dataset 3: 11,796 agents, 21,592 persons, 175 organizations). Here, it seems that these mappings of agents do not depend on the mapper but on the provided data.

4. Linked Data References vs. Literal Statements

Broadly speaking, an RDF statement can have either a literal (a possibly typed string) or a reference to a resource (an IRI, a blank node or an RDF container type). Since the DM2E model strongly recommends using literals and IRI exclusively, the relationship between statements referring to literals or resources and the total number of statements in a dataset reveals differences in the datasets as can be seen in figure 3. When the datasets are grouped by the percentage of

[8] Internationalized resource identifier. An extension of URI allowing unencoded Unicode characters in most places of a URI (RFC 3987).

literal statements, clusters of similar percentages appear according to the respective MI - independent of the metadata content.

For example, the percentage of literal statements in DP V-Dataset 7 (28.273%) and DP VI-Dataset 8 (28.270%) is almost equal, yet the content is vastly different (collection of digitized prints of various genres and ages vs. personal correspondence of an 19th century scholar), the metadata originally created by different data providers (research project vs. library) and in different formats (TEI vs. EAD). The only commonality between the datasets is that the same organization (MI D) created the mappings to DM2E. Therefore, we put forth the correlation between the ratio of literal statements and the mapping institution is much stronger than between ratio of literal statements and similarity of the original data.

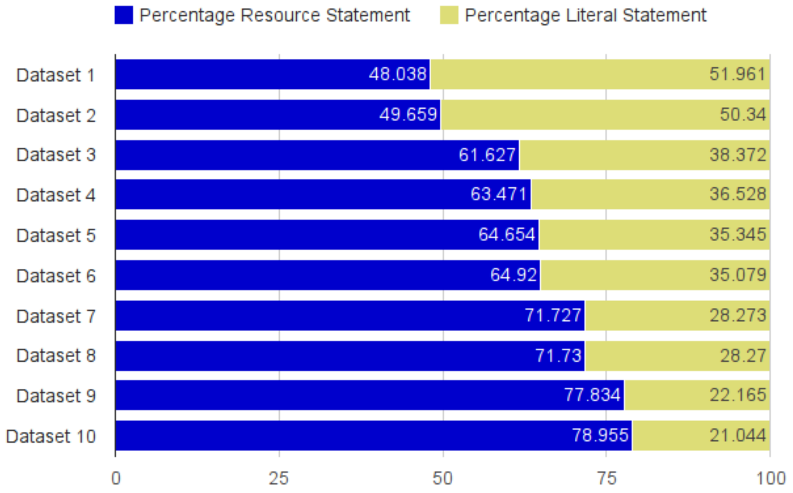


FIG. 3: Ratio of statements with literal statements to resource statements per dataset.

While the relationship between resource and literal statements gives some insight into how MIs structure the data, it does not answer questions pertaining to the quality and usefulness of literal statements. To tackle this problem, the literal statements containing properties with literals allowed as their range were clustered into three groups (see fig. 4). The "preferred" literal statements (properties that are either mandatory, recommended or increase the descriptive content)[9] are a sign of data quality since they enhance the descriptiveness of the data, improve the search and browse experience and granularize textual information. The "neutral" literal statements are those neither preferred nor unwanted, i.e. properties where it is not important for contextual information if they refer to resources or literals. Lastly, the "deprecated" literal statements are statements with those properties that allow both literals and resources in their range, yet the data providers chose to use literals.[10] Even though the label implies it, it is not necessarily a wrong choice to use literals when they are allowed as an alternative to an IRI. However, inconsistent usage is detrimental to the homogeneity of the data, requiring data consumers to use more complex queries to capture both types of statements and are often a sign for poor structure within the data.

As can be seen in figure 4, there is some evidence that the relationship of the number of preferred and deprecated literal statements is correlative with the mapping institution. For

[9] Preferred properties in literal statements: *skos:prefLabel*, *rdfs:label*, *skos:altLabel*, *dc:description*, *dm2e:displayLevel*, *edm:type*, *dc:title*, *dm2e:subtitle*, *dc:language*, *dc:format*, *dc:identifier*.

[10] "Deprecated" properties in literal statements: *dc:rights*, *dcterms:created*, *dcterms:modified*, *dcterms:issued*, *dcterms:temporal*, *rdaGr2:dateOfBirth*, *rdaGr2:dateOfDeath*, *rdaGr2:dateOfEstablishment*, *rdaGr2:dateOfTermination*. The model recommends for time-related properties the use of *edm:TimeSpan* resources but also allows *xsd:dateTime*/*xsd:gYear* or *rdf:Literal*.

example, the data produced with mappings by MI A (Dataset 1 and 2) and MI C (Dataset 5 and 6) is very coherent in this regard. However, for the datasets produced by MI B (Dataset 3 and 4) we see a slight variance, for the datasets produced by MI D (Dataset 7 and 8) even a significant variance in the ratios. Taking the more specific grouping into account, the preferred-deprecated ratio is much more influenced by the original metadata than the overall literal-resource ratio. Considering the data produced by MI D, it is remarkable that the one dataset (Dataset 7) contains the largest proportion of deprecated literal statements within the set of datasets, whereas the other dataset (Dataset 8) contains no deprecated statements at all.

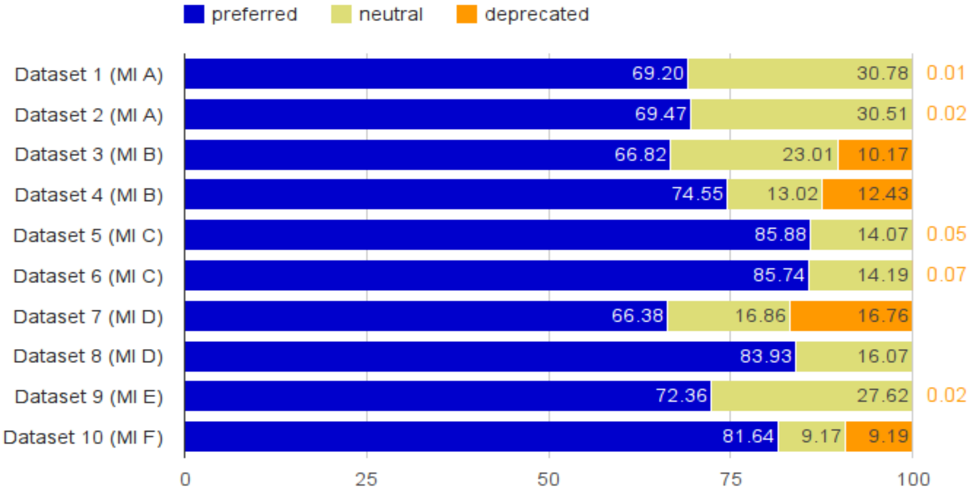


FIG. 4: Distribution of "preferred", "neutral" and "deprecated" literal statements within the datasets.

5. Variance of Statements and Redundancy of Data in Triples

To measure the redundancy of data in triples, we introduce the measure of Predicate-Object-Equality-Ratio (POER-*n*), which is defined as the percentage of triples that share the same predicate and object with at least *n* other statements. In other words, POER-*n* measures how many statements state the same facts about different subjects. The smallest possible POER-*n* of the datasets in DM2E, POER-1, ranges from 0.08% (Dataset 5) to 2.48% (Dataset 3). While impressive as a signifier of structural redundancy, using POER-*n* to assess data-intrinsic redundancy proves to be much more difficult. First of all, there is a lot of duplication required by the triple structure of RDF, i.e. *rdf:type* statements have a limited range of possible values defined by the DM2E model. Certain literal properties have even smaller ranges. Other areas of redundancy can be explained by the original metadata, such as manuscripts being published in the same year or by the same author. Some redundancies, however, can point to problems. For example, redundancies in *dc:subject* statements will, when passing a certain frequency threshold, not be discriminatory for any kind of search (e.g. assigning the keyword "philosophy" to any CHO). Redundant *dc:title* statements can show mapping errors or missing content. For example, if many *dc:title* statements contain the text "Untitled Page" or just a page number, the content may have been mapped incorrectly.

Hence, the usefulness of POER-*n* is very dependent on the value of *n*. Whereas the bulk of the statements contained in POER-1 or even POER-100 can be discarded as arbitrary similarities, a high POER-1000 or POER-10000 cannot be easily explained with random chance. If the same fact is stated about 10,000 different subjects within a dataset, this is a strong indicator that either the original metadata is very homogenic (e.g. by the same author or released in the same year) or that the data is not properly internally aligned (e.g. hundreds of different auto-generated *skos:Concept*s with the same *skos:prefLabel*). Instead of setting *n* to an arbitrary number, a lot can be gained by using the number of instances of certain classes as the threshold, for example, in the case of DM2E, the number of *ore:Aggregation*/*edm:ProvidedCHO* instances. The exact mechanics of how to fine-tune POER-*n*, finding proper threshold values and visualizing both, the POER-*n* and the statements it represents, is still subject to further research.

Figure 5 presents the average number of statements per instance of a class within a dataset (ANOS). We see that the data mapped by MI C is very homogenous with regards to the ANOS, for both *ore:Aggregation*/*edm:ProvidedCHO* and contextual classes. Obviously, the workflow for the RDFization of the original data used by MI C is organized in such a way (e.g. by reusing the same XSLT scripts) that the resulting RDF follows a relatively rigid structure.

For the *edm:ProvidedCHO* instances, we see a significant higher ANOS for data mapped by MI D. Since the data is generated from very different input formats, the deciding factor here is apparently MI D's thorough mapping process, producing more statements by normalizing unstructured fields, adding alternative titles, different languages etc.

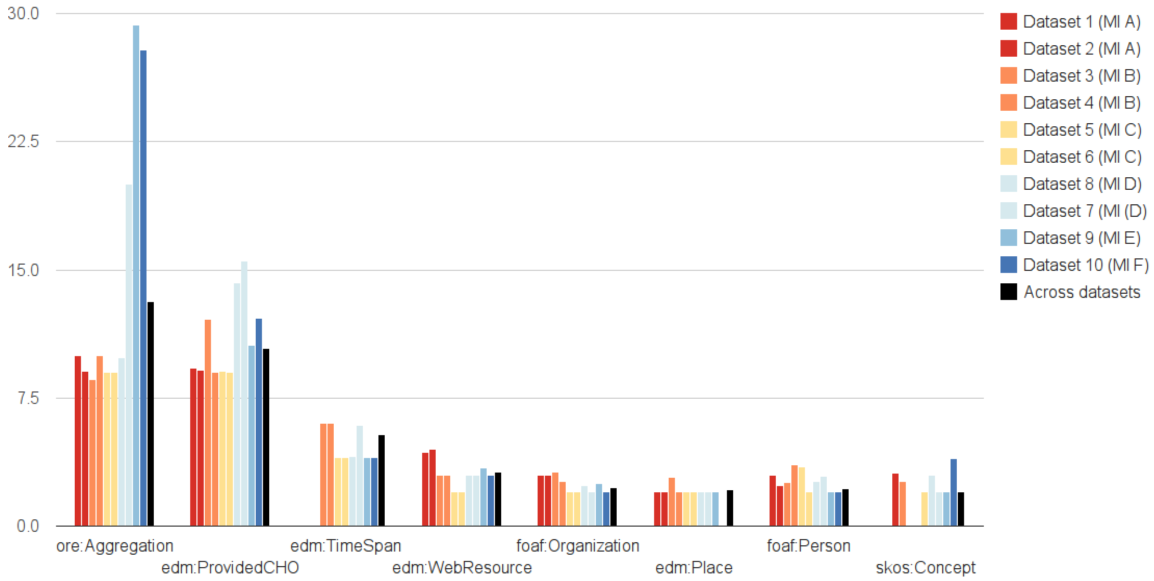


FIG. 5: Average number of statements per class per dataset.

The three outliers with significantly more-than-average ANOS for *ore:Aggregation*s are all generated from TEI data. Apparently, TEI's exhaustive mechanics for adding metadata to the header of a TEI document heavily and positively influences the richness of the metadata on aggregation level. While still slightly above average, the ANOS for *edm:ProvidedCHO* from TEI data is much lower than for *ore:Aggregation*, leading to the conclusion that TEI is a top-heavy format, inciting TEI producers to create exhaustive meta-metadata describing the provenance of the TEI document rather than the manuscript itself.

Looking at the distribution of ANOS for *edm:WebResource* instances, clusters of very similar ANOS defined by the respective MI emerge. The explanation for this is that most information assigned to *edm:WebResource* instances is boilerplate (format and rights information mostly) with only the IRI of the *edm:WebResource* instance itself changing.

In general, the distribution of ANOS across datasets is more homogenous for contextual classes (*foaf:Person*, *foaf:Organization*, *edm:Place*, *edm:TimeSpan*, *skos:Concept*) than for manuscript-related classes (*ore:Aggregation*, *edm:ProvidedCHO*). The main reason for this is that ANOS for the former is significantly smaller than for the latter, i.e. relatively few statements are asserted about instances of contextual classes (the highest ANOS for contextual classes is 3.96 for *skos:Concept* in Dataset 10). On the other hand, this is also a sign that there is still potential for possible improvement on account that, e.g. digitization projects focusing on the written legacies of individuals tend to have extensive dossiers about the context (like places, persons and concepts). Apparently, the full richness of this data is not yet fully ported over to the RDFized data.

6. Being Linked Open Data - Usage of different Ontologies

The Linked Data principles recommend using existing namespaces and ontologies. The DM2E model included a number of other ontologies and encouraged data providers to map their data using properties from them. Figure 6 shows the ontologies and their number of properties referenced by the DM2E model as well the number of properties used by data providers.

Every ontology is used, however, not all properties are used: of DM2E, slightly more than 50% of the offered properties are used, around 66% of EDM. Most of the properties of the DC and BIBO ontologies are used (75%). Vocabularies like DC and DCTerms have fewer resources in the model than DM2E but they are more often used. Other ontologies like rdaGr2 provide very specific properties for very specific contextual classes which are also often not mapped (e.g. the already mentioned *rdaGr2:dateOfEstablishment*). Even though the two CIDOC-CRM properties in the model, *crm:P79F.beginning_is_qualified_by* and *crm:P80F.end_is_qualified_by*, are also very specific, they serve an important case: they are used to indicate how accurate a timespan is.

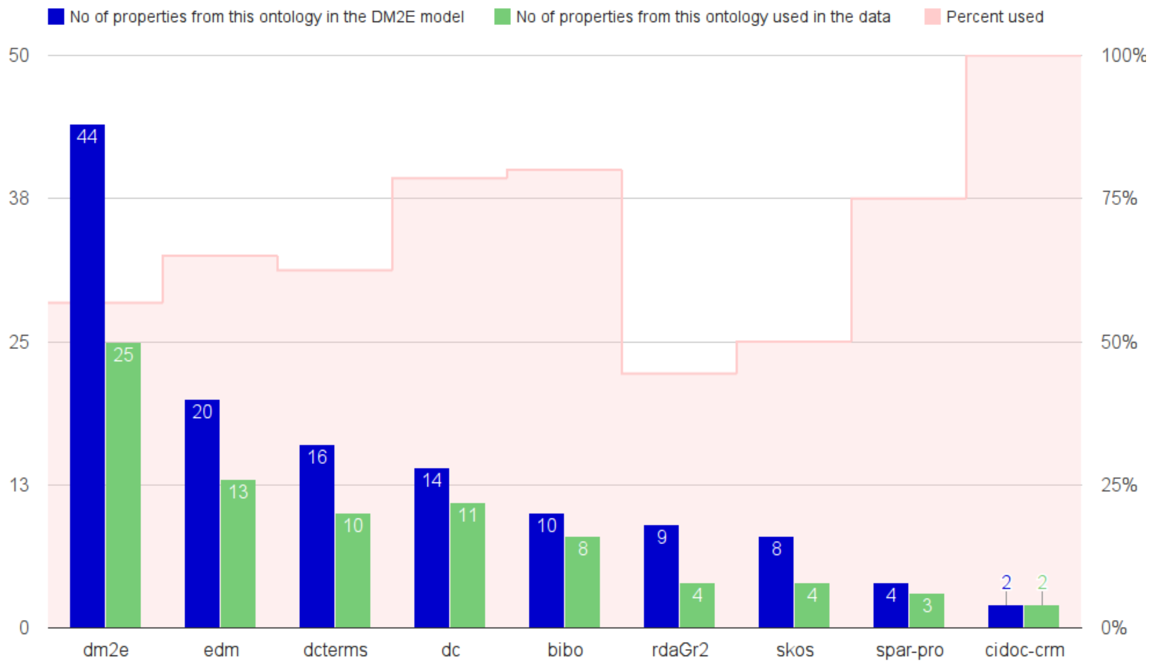


FIG. 6: Number of properties defined in the DM2E model vs. number of properties actually used in the data, by referenced ontology.

The fact that only half the properties defined in the DM2E model are actually used (see also fig. 1) deserves closer scrutiny, however. Because the ontology is being developed by DM2E for DM2E, this cannot be explained with the specificity of the domain of the ontology, but with the dynamics of the process of ontology development: In the early stages, the intricate knowledge of data providers about the details of their data led them to require increasingly semantically narrow properties from the DM2E ontology engineers (e.g. *dm2e:honoree* or *dm2e:wasStudiedBy*). However, when the MI (which do not necessarily coincide with the DP, see table 1) started implementing the mappings, many of those requirements were dropped due to the specific properties being hard to map or not being readily discernible from the original metadata. Over the course of many cycles of mapping, data ingestion and refinement of the data model, new properties have been added but unused properties were never dropped.

7. Conclusion: Linked Data Mapping Cultures

The analyses have shown that the particular mapping institution plays an important role in the way that data actually is represented after a mapping process. Datasets mapped by the same MIs have similar characteristics in the various analyzed aspects, e.g. which resources are used for the mappings and which are not. The representation of the data before the mapping has a less significant influence on the structure of the mapped data as has the domain or CHO types. The source format is reflected in the number of provided statements, e.g. whenever TEI is used (where the full text of an object is also annotated and can be used for mappings), many more statements are produced.

As already identified in previous model evaluations, mapping institutions do not make use of the full range of possible ontology elements that could be mapped. Models, including the DM2E model, could be reduced (especially when only a small percentage of specific vocabularies is used as shown in the last figure). Contextual resources are not mapped as thoroughly as the core classes for the representation of the object (*edm:ProvidedCHO*) and its metadata record (*ore:Aggregation*).

From a user's perspective, the Linked Data representation should be derived from the source data by a function of the source data and not strongly be influenced by the specifics of the mapping process. While technical means such as the quantitative analyses presented here help make the skew more evident, it can eventually only be rectified by a more agile development process that involves all stakeholders balancing semantic expressivity with data interoperability, peer-review of mappings or ongoing evaluation of mappings and mapped data, improved and extended mapping guidelines with a strong focus on reusability and sustainability of data and data model. From a Linked Data mapping cultural perspective, our conclusion is that ontologies should not just be extended to fit new requirements but also pruned from over-specific bloat regularly and that this can only be achieved when ontologists, data providers, mapping institutions, developers and data consumers incessantly communicate, compromising between semantic accuracy and technical feasibility.

Acknowledgements

This work was supported by a grant from the European Commission "ICT Policy Support Programmes" provided for the DM2E project (GA no. 297274). The authors would like to thank all colleagues of DM2E as well as the European Commission for their support.

References

Alexander, Keith, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. (2009). Describing Linked Datasets. On the Design and Usage of VoID, the "Vocabulary of Interlinked Datasets". In Bizer et al. (Eds.), Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings. Retrieved, May 14, 2014, from http://ceur-ws.org/Vol-538/.

Auer, Sören, Jan Demter, Michael Martin, and Jens Lehmann. (2012). LODStats – An Extensible Framework for High-Performance Dataset Analytics. In ten Teije et al. (Eds.), Knowledge Engineering and Knowledge Management. 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012, Proceedings (pp. 356-362). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-33876-2.

Carroll, J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. (2005). Named Graphs. In Journal of Web Semantics, 3, 247-267.

Dröge, Evelyn, Julia Iwanowa, and Steffen Hennicke. (2014a). A specialisation of the Europeana Data Model for the representation of manuscripts: The DM2E model. In Libraries in the Digital Age (LIDA) Proceedings, Volume 13, 2014. Retrieved, July, 24, 2014, from http://ozk.unizd.hr/proceedings/index.php/lida/article/view/117.

Dröge, Evelyn, Julia Iwanowa, Steffen Hennicke and Kai Eckert. (2014b, March). DM2E Model V1.1 Retrieved, May 12, 2014, from http://pro.europeana.eu/documents/1044284/0/DM2E+Model+V+1.1+Specification.

Europeana Data Model Primer, v14/07/2013. (2013, July). Retrieved from: Europeana Professional website. Retrieved, April 28, 2014, from http://pro.europeana.eu/ documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.

Heath, Tom, and Christian Bizer. (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology (Vol. 1). Morgan & Claypool.

Klimek, Jakub, Jirí Helmich, and Martin Necasky. (2014). An analysis supported by numerous visualizations Application of the Linked Data Visualization Model on Real World Data from the Czech LOD Cloud. Linked Data on the Web (LDOW 2014) Workshop. Retrieved, May 14, 2014, from http://events.linkeddata.org/ldow2014/papers/ldow2014_paper_13.pdf.

Palavitsinis, Nikos, Nikos Manouselis, and Salvador Sanchez-Alonso. (2014). Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process. Journal of the Association for Information Science and Technology. doi: 10.1002/asi.23045.

Seiffert, Florian. (2001). Eine Analyse der Verbunddaten des HBZ. ABI-technik 21(2): 125-146.

Smith-Yoshimura, Karen, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlison de Ortiz, Hugh Taylor. (2010, March). Implications of MARC Tag Usage on Library Metadata Practices, OCLC Online Computer Library Center, Inc. Retrieved, May 14, 2014, from http://www.oclc.org/research/publications/library/2010/2010-06.pdf.

The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation

Mark A. Matienzo
Digital Public Library of America, USA
mark@dp.la

Amy Rudersdorf
Digital Public Library of America, USA
amy@dp.la

Abstract

The Digital Public Library of America (DPLA) aggregates metadata for cultural heritage materials from 20 direct partners, or Hubs, across the United States. While the initial build-out of the DPLA's infrastructure used a lightweight ingestion system that was ultimately pushed into production, a year's experience has allowed DPLA and its partners to identify limitations to that system, the quality and scalability of metadata remediation and enhancement possible, and areas for collaboration and leadership across the partnership. Although improved infrastructure is needed to support aggregation at this scale and complexity, ultimately DPLA needs to balance responsibilities across the partnership and establish a strong community that shares ownership of the aggregation process.

Keywords: metadata aggregation; metadata remediation; harvesting; software development; community development; JSON-LD

1. Introduction

The Digital Public Library of America (DPLA) recently celebrated its first anniversary aggregating the riches of America's libraries, archives, and museums and sharing them through a single portal. With its Hubs (the 20 direct partners from whom DPLA harvests records) and their partners (approximately 1,300 in all), DPLA has worked to make these resources freely available to the world. After a year focusing resources on growth, with the DPLA holdings more than tripling to over seven million records in twelve months, it seems an appropriate time to take stock of the technologies and processes within which this work occurs, as well as the data models used to aggregate the Hubs' various metadata standards and the nature of collaboration between DPLA and the Hubs. It is important to identify areas both of success and improvement that have become apparent since the launch in April 2013. This assessment takes into consideration outside variables, as well, including feedback from Hubs, users of DPLA's open and freely available application programming interface (API), and others interested in the DPLA technology stack and metadata model. A few areas of future work have been identified, which will help to create a roadmap for ongoing investigation and development. It is hoped, too, that this process will involve current and future partners, and create a community of practice around these open source technologies and metadata management systems.

2. Development, implementation, and current status of DPLA infrastructure

DPLA launched its services on April 18, 2013, with 2.4 million records from 16 Hubs (and their over 900 partners) after a two-year planning phase. The components that make up the technology stack that supports the infrastructure are lightweight and open source, which allowed DPLA's initial technical implementation team to prototype and deploy working iterations quickly. DPLA also developed a metadata application profile, or MAP (Digital Public Library of America, 2014a), based on existing data standards and models. In addition to the ingestion system described below, DPLA's infrastructure also provides both an application programming interface (API) and a public user interface that serves as the primary discovery system for the ingested metadata. The platform, or API layer, is a Ruby on Rails web application that provides an abstraction mechanism over the primary data store and search index. The portal, or user-facing front-end application, is built on Ruby on Rails, and is a client of the platform application.

The DPLA technical infrastructure was implemented over a period of 18 months, which demanded a relatively short build-out process. During the initial implementation period (October 2012-April 2013), the DPLA Assistant Director for Content undertook primary responsibility for developing the metadata mappings, and a team of contractors developed the metadata ingestion system and other areas of infrastructure and ran the ingestion processes. Since late 2013, the DPLA staff has steadily grown, including the hiring of a Director of Technology (December 2013), two Technology Specialists (January and May 2014), a Data Services Coordinator (August 2014), and a Metadata and Platform Architect (August 2014). During this time, DPLA has undertaken most of the responsibility for maintaining the existing infrastructure, overseeing the ingestion process, and identifying areas for improvement.

2.1 The DPLA Metadata Application Profile

The DPLA Metadata Application Profile (MAP) is an extension of the Europeana Data Model, or EDM (Europeana, 2014). Version 3, the first public version of the MAP, was developed in early 2013 by DPLA staff and others, in collaboration with Europeana staff and public data specialists who provided input during an open review period in late 2012. Like EDM, the MAP incorporates or references a variety of standards and models, including the Dublin Core Metadata Element Set, Dublin Core Terms, the DCMI Type Vocabulary, OAI-Object Reuse and Exchange, and others. While based on EDM, the DPLA MAP nonetheless slightly diverges from it. First, one of the MAP's core classes, the Source Resource (dpla:SourceResource), is defined as a subclass of the corresponding class in EDM (Provided Cultural Heritage Object, or edm:ProvidedCHO). The primary motivation for this was to make clear that the properties of dpla:SourceResource in some cases may have different cardinalities or requirements than those defined for edm:ProvidedCHO. In addition, because of limitations on both the data available from DPLA's providers and the geocoding enrichments implemented near launch, DPLA developed its own spatial location class, dpla:Place.

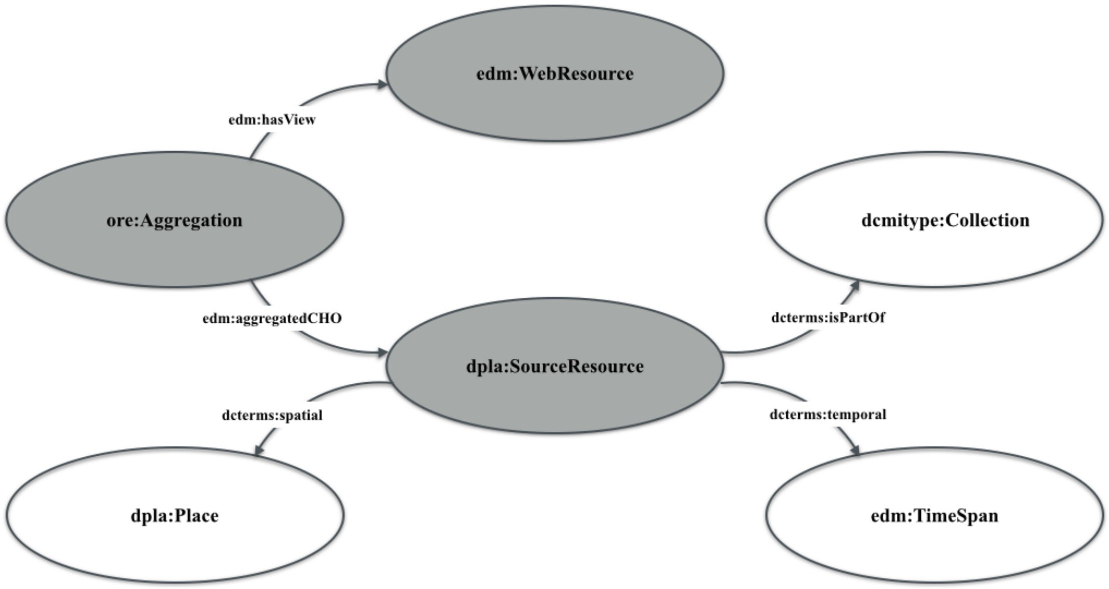


FIG. 1. Core classes and relationships in the DPLA Metadata Application Profile, versions 3 and 3.1.

DPLA staff reviewed and revised the requirements for the MAP in mid-2014, and released MAP version 3.1 in July 2014. Many of the differences between MAP versions 3 and 3.1 relate to cardinality requirements, which were changed based on recognition of the properties DPLA could not reliably receive, map, or otherwise derive from metadata provided by Hubs. DPLA also added a new property (Intermediate Provider, or dpla:intermediateProvider) to allow for the declaration of an entity understood to be distinct from the two provider-related properties within EDM (edm:Provider and edm:dataProvider). MAP version 3.1 defines an Intermediate Provider as "an intermediate organization that selects, collates, or curates data from [an edm:dataProvider] that is then aggregated by [an edm:Provider] from which DPLA harvests" (Digital Public Library of America, 2014a). Beyond these changes, MAP version 3.1 also contains several changes which bring it towards further alignment with EDM, such as clearly identifying the super-properties for a given property when available, aligning internal properties with EDM definitions, adding the edm:hasType property to express genre statements, and adding the edm:rights property. The addition of edm:rights allows for association of rights information available at from a given URI to two core classes within the MAP.

2.2 Ingestion system and workflow

The DPLA ingestion system (Digital Public Library of America, 2014b) is an application, written in Python using the Akara (2010) framework, that provides REST endpoints for web services to transform or enrich data serialized in JSON. The primary DPLA data store is a BigCouch/CouchDB document-oriented database, with metadata both stored and serialized using JSON-LD 1.0 (Sporny, Kellogg, and Lanthaler, 2014). Once stored in BigCouch, all ingested metadata is indexed using Elasticsearch, a REST-based search server built upon Apache Lucene. Additional scripts that support or control the ingestion process are also written in Python. The ingestion workflow for a given *ingestion source* has a designated *ingestion profile*. In most cases, Hubs only provide one ingestion source, but a small number of Hubs are continuing to develop internal systems to support the single-ingestion-source model that is, technically, a requirement to DPLA participation. Accordingly, a single Hub that has more than one ingestion source may have multiple ingestion profiles. Each ingestion profile is a JSON document containing configuration information such as the type of harvest, (e.g., OAI-PMH, site-specific API, static files, etc.), location of an HTTP endpoint if applicable (e.g., the OAI-PMH provider URI), the specific mapping and enrichments to be applied, and other internal settings required by the ingestion system.

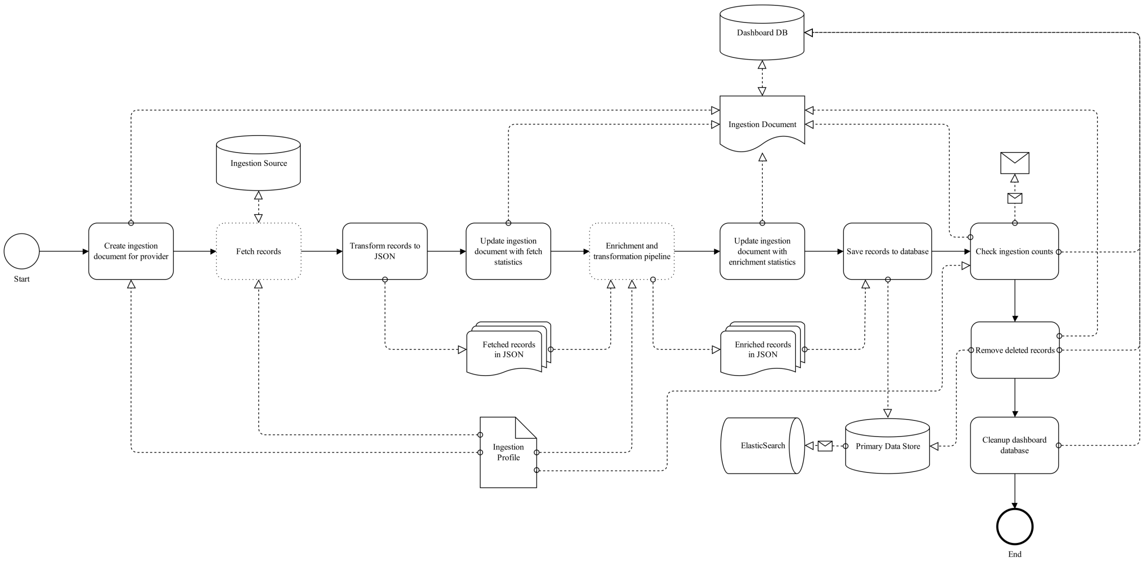


FIG 2. Overview of the DPLA ingestion workflow.

The ingestion workflow is invoked by a support script that reads the ingestion profile for a given source and creates an *ingestion document* in the *dashboard database* for a given *ingestion process*. The ingestion document contains data about the state of particular *ingestion task* (e.g., whether a specific step has started, completed, or failed). The dashboard database also temporarily contains a representation of each fetched record to allow staff to identify what parts of an ingested record have changed. Once the ingestion document is created, the staff running the ingestion process invokes the *fetch task*, which obtains the metadata to be ingested from the source defined in the profile. The metadata is then deserialized from its native format (typically XML), reserialized as a JSON expression of the original data, and persisted to disk in a temporary location. Once the fetch process is complete, the ingestion document is updated to contain the location of the data transformed to JSON.

The ingestion staff then invokes the *transformation and enrichment tasks*. These tasks map and transform the JSON-serialized metadata to the DPLA MAP, and normalize, enhance, and augment the metadata using a "pipeline" that orchestrates requests to the application's REST endpoints (see section 2.3 for more information). Once complete, the records are temporarily persisted to disk as a JSON-LD serialization of the MAP, and the ingestion document is updated with information about transformation and enrichment processes, including location of the transformed records and the extent of any failures within the process. The ingestion staff then runs the *save task*, which reads the MAP-compliant JSON-LD records and persists them to the primary data store. After the save process completes, the ingestion staff runs the *check ingestion counts task*, which identifies the number of new, updated, or deleted records for each ingestion process and automatically alerts the identified staff when those values are above a certain threshold defined in the ingestion profile. Finally, the ingestion staff runs two concluding tasks: the *remove deleted records task* and the *dashboard database cleanup task*. Both tasks remove objects from the primary data store or dashboard database. These objects correspond to the metadata from ingested records that were either deleted from the ingestion source by the provider (e.g., as identifiable using the <deleted> element from an OAI-PMH provider) or otherwise not present or available during a given ingestion process.

2.3 The metadata transformation and enrichment pipeline

Most of the work to transform, normalize, and enhance the metadata ingested into DPLA occurs as part of the *transformation and enrichment pipeline*, which executes a list of specific steps defined in an ingestion profile in a specific, linear order. Each of the steps is implemented in the ingestion system as a module mounted at a defined REST endpoint. Each of the endpoints receives JSON data over an HTTP POST request, and returns JSON data, either modified if the step was applicable and successful or unchanged if the step was inapplicable or if it failed. Most of the ingestion profiles share a number of common steps, and the modular design allows DPLA to easily reuse them and add extra parameters as needed.

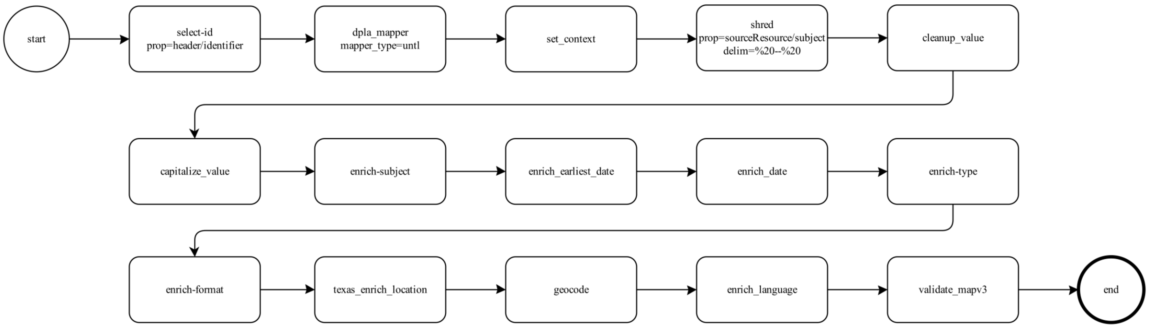


FIG 3. Sample transformation and enrichment pipeline for ingestion from the Portal to Texas History.

At a minimum, the pipeline must contain two steps: one that selects the source of the identifier from the ingested metadata (which is required for persistence), and another that transforms and maps the metadata to the DPLA MAP. Despite the pressures related to launch, DPLA was also able to implement some degree of normalization and enrichment. Much of the DPLA staff's ongoing work involves revising and ensuring that these normalization and enrichment modules remain robust and error-free. At a minimum, the enhancements applied to most metadata ingested into DPLA include what Hillmann, Dushay, and Phipps (2004) term "safe transforms," through global cleanup of values to address minor differences in capitalization, punctuation, or whitespace, or alignment and reconciliation of terms against comparatively small controlled vocabularies such as the DCMI Type Vocabulary or ISO 639-3 language codes. In addition, the ingestion system undertakes more complex transformations based on diversity of practice, such as normalizing dates or date ranges to a common format, and "shredding" a string literal based on a given delimiter to yield multiple values. In addition, the ingestion system also includes a geocoding enrichment service, which uses external services to take geographic name values and geocode them to return latitude and longitude pairs, and then uses those coordinates to build out a geographic hierarchy. More details about these services are provided below.

The quick lead up to the launch meant turnaround times were limited and the need to ingest metadata created using different schemas under varying practice and assumptions meant that some areas of work on the transformation and enrichment pipeline had to be reprioritized. Work during the initial ingest, which took place roughly between February and mid-April 2013, focused on mapping and the conceptual alignment of fields from the initial 16 Hubs, rather than on the review and quality control of the actual values. Likewise, a loosening of validation against the MAP assertions was necessary to ensure that goals and timelines were met. This period focused on return on investment in the strictest sense: providing the best data in the shortest period of time with the least remediation. In addition, since MAP version 3 was only finalized approximately three months before launch (and only days before the first ingests began), additional changes to the ingestion code and DPLA's Platform API were necessary to ensure that all of the data was available through the portal by mid-April 2013.

3. Concerns and challenges

The technology and data model established for the launch has served DPLA well. It has effectively aggregated over seven million records, enabling hundreds of users to utilize the API and effectively build apps, and more than a million users to search and enjoy the resources available through the portal. With sustained use and the ongoing need to continue the ingestion of metadata from both current and future Hubs, challenges have arisen that signal a need to consider potential new options for aggregation, storage, and delivery.

3.1. The ingestion process

Ingest remains a very hands-on endeavor. Once a Hub's data is mapped to the DPLA Metadata Application Profile (by the Assistant Director for Content, at the time of publication), a new ingestion profile is written (by DPLA technology staff) that delineates the harvesting, transformation and enrichment steps. In addition, despite using common metadata standards (e.g., DCMES or MODS) or harvesting protocols (e.g., OAI-PMH), differences in local implementation often require DPLA technology staff to modify or supplement implemented mappings, employ new transformation services, or resolve other inconsistencies before an ingest moves to production. For example, several Hubs have found it difficult to reliably provide URIs for thumbnail images for the items associated with the metadata ingested by DPLA. As this information is mandatory in MAP version 3.1, DPLA technology staff must often undertake a degree of reverse engineering to add an enrichment step that identifies or constructs this URI. Nonetheless, while discussions between Hub and DPLA personnel lead to good results, the process of getting a new data set into production often lasts between four and eight weeks.

The ingestion process itself is also resource intensive, and as described above, the architectural paradigm of the current ingestion system currently expects that a consistent transformation and enrichment pipeline be used across *all* ingestion processes from a given ingestion source. A large number of processes are applied to all incoming ingests regardless of the metadata schema used or quality of the metadata received. Currently, data from each Hub is reingested *in its entirety* monthly, every other month, or quarterly, depending on the frequency of local updates. Accordingly, each step defined in the transformation and enrichment pipeline runs during each ingestion process. This ultimately leads to the potential for some enhancements to be lost or misapplied if a Hub has modified its metadata in the interim. Improved control over the enrichment workflow, such as enabling or disabling certain processes for a scheduled ingestion process for a specific Hub, and supplementing those enrichments with provenance information, could provide better control and reduce complexity of ingestion on an ongoing basis. And while the process has been internally standardized, it remains somewhat opaque to some Hubs, especially those who may not be familiar with the languages in which the transformation and enrichment pipeline modules are written. In the experience of DPLA, this also points to the need for improved unit tests and documentation that make the intent of the pipeline modules clearer to domain experts without programming knowledge.

Other challenges to the current model that have come to light over the past year include the inconsistency of some of the enrichment and normalization processes that are applied to all collections. For example, DPLA staff recently identified that structured spatial information (i.e., a place hierarchy) provided by some Hubs was not successfully mapped to the property required for the literals to appear in the user interface (skos:prefLabel). Diagnosis of issues in the enrichment process proves to be an ongoing challenge for DPLA given that the ingestion system does not track the provenance of statements created or modified during transformation and enrichment. In addition, while the DPLA MAP is a data model based upon RDF, the current infrastructure has not yet implemented a complete expression of the constraints defined by it. These limitations originate mostly because the current implementation of validation relies on a simplified expression of the MAP using JSON Schema (Galiegue, Zyp, and Court, 2013), with any validation of the statements about a given item against the MAP currently limited to cardinality checks and simple controlled value verification based on the JSON serialization of the data.

Another area in which DPLA continues to face challenges is the geocoding enrichment process, which retrieves a "best guess" set of coordinates for a term from the Bing Maps API, and uses those coordinates to build out the rest of a geographic hierarchy for that term using the Geonames API. For the value "Charlotte (NC)," the values "35.226944, -80.843333" are automatically assigned via the Bing Maps API to indicate the geographic center of the city. Then, those coordinates are sent to the Geonames API to extract the geographic hierarchy for Charlotte, i.e., United States -- North Carolina -- Mecklenberg County -- Charlotte. This is rich and valuable data that allows DPLA to plot "Charlotte (NC)" on the interactive map in the portal. Like any scaled transformation, this process is not fail-safe, as a careful study of the map exposes. For example, consider a record with the spatial value of "Wisconsin." In this model, the coordinates for the central point of the state identify a hierarchy that contains county-level information (United States -- Wisconsin -- Portage County), which introduces data that can be misleading, if not erroneous. In addition, DPLA staff has discovered that external web services like the Bing Maps API often update the data they provide or their indexing mechanism, which has led to inconsistencies in the geocoding enrichment processes over time. Considering the lack of confidence about the geocoding process and the inability to track provenance of statements in DPLA's current infrastructure, DPLA has chosen not to implement reconciliation of geographic names with URIs from sources such as Geonames until these issues can be addressed.

3.2. The metadata

Over the past year, DPLA staff has had the opportunity to work closely with Hubs from across the United States. Not surprisingly, the Hubs employ various metadata standards, maintain data in many different repository types, and manage localized workflow models. The process of aggregation, and especially enrichment and normalization, has been eye-opening for most of the parties involved. DPLA staff knew even before harvesting began in early 2013 that the process would be complex and not without challenges, as evidenced by past work on projects such as the National Science Digital Library (Lagoze et al., 2006), the Digital Library Federation Aquifer Project (Riley et al., 2008), and Europeana. One immediate revelation was somewhat surprising, however. The greatest difference between collections—and the source of the most difficulties—is not the metadata schemas employed or repositories used, but the extent to which simple metadata, like unqualified Dublin Core exposed over OAI-PMH, must be processed, and, more importantly, how metadata is input and managed locally.

When data is shared in MODS, MARCXML, or even qualified Dublin Core, the richness and completeness of the records transfers relatively easily to the DPLA model. Not surprising, of course, is that the more granular the original record, the better the output at the other end. However, unqualified Dublin Core—most often exposed over OAI-PMH—requires a great deal more analysis and a greater number of complex transformations to identify and map discrete values in a single field to multiple fields in the MAP. For example, specific transformation and enrichment modules are created to determine when a dc:coverage field contains only spatial information, spatial information together with temporal information, or only temporal information. Similar issues, although no less challenging, arise from the varied interpretation of values in dc:source, dc:contributor, dc:relation, dc:type, and others. In evaluating the importance or the efficacy of these transforms, DPLA is reminded that "minimally descriptive metadata … is still minimally descriptive after multiple quality repairs" (Lagoze, et al. 2006). In some ways, this problem is exacerbated further given that Hubs are often aggregators themselves. The degree to which values have been "dumbed down" is not always well documented in terms of how or where this simplification occurred.

It also became immediately clear when a Hub, or its partners, consistently employed and applied (or didn't) controlled vocabularies. While most Hubs follow general guidelines for geographic names (e.g., selecting terms from vocabularies like TGN or LCSH), they are not always applied consistently. Again, this is in part because many Hubs are themselves aggregators of content from hundreds of partners. On DPLA's long-term roadmap for implementation is the work to implement reconciliation of string literals against large controlled vocabularies. Interestingly, in many collections, Hubs' partner names are not taken from controlled vocabularies, or if they are, either this is not indicated in the data or the authorized form of name lacks important contextual information. This has led to a surprising number of errors or unfamiliar values in the data, at least initially. One Hub utilizes the Library of Congress Name Authority File to create their controlled list of partner names. While on the surface this seems like a prudent approach, until the terms are associated with URIs and are augmented with more information, many of the names have very little meaning outside of their local context. For example, not everyone can readily associate the LC Name Authority "J. Y. Joyner Library" with East Carolina University (the parent institution).

4. Responses and requests from DPLA Hubs

DPLA personnel have actively worked in partnership with Hubs to identify and openly communicate quality issues in the data that they are sharing. Hubs have been responsive and often eager to make updates and changes to data and even the mappings in their local systems to better align with international practice and the DPLA data model. All agree that this has meant better data quality at both the local and global level. Through this process, Hubs have shared thoughts on ways that ingest could be improved. In some cases, they have begun local development on tools that transform and enrich their data before it reaches DPLA. Some of the requests DPLA has heard align well with its own internal priorities and needs.

4.1. Greater control over and feedback during the ingestion process

As mentioned earlier, the community feels strongly that they would benefit from an "ingestion dashboard" that offers a selection of enrichment processes from which Hubs could choose to apply to their data during the ingest process. Because the Hubs know their data best, enabling access to an ingestion dashboard and involving them as early as possible in the initial mapping process would give the Hubs more control over the way their data is exposed via DPLA. Also, it would shed light on what remains a somewhat opaque process for those who are not proficient with the technologies in use. In the interim, DPLA has developed a basic content quality assurance dashboard for internal use and review by Hubs before an initial ingestion reaches the DPLA production data store. The dashboard application is part of the platform API infrastructure, and provides a stripped-down user interface for search and browse of ingested metadata, and the generation of reports on metadata output from the transformation and enrichment pipeline. In addition, integrating tools that provide better visual representations of how metadata is mapped at ingestion and presented in the DPLA portal interface (e.g., Gregory and Williams, 2014) would benefit stakeholders across the DPLA network.

4.2. Access to data quality reports

As part of the initial ingestion process for a new Hub, a series of reports are produced that enable DPLA staff to review the values in each field mapped to the DPLA application profile. For each property, two reports are produced: an itemized list of all values in the field and the corresponding DPLA record identifier, and a count of all of the values in that field. The reports are produced from the enriched data, after geocoding and normalization have been applied. Some Hubs, especially those with repository systems that cannot easily generate aggregated reports for a given element or predicate, have requested access to reports on their *unprocessed* data. This would allow them to assess their metadata and perform remediation locally, before it is ever harvested by DPLA. While valuable, this will require significant re-engineering of the ingestion system before it can be implemented.

4.3. Upstream data flow: receiving DPLA-provided enrichments

The greatest challenge, but one that several Hubs have voiced interest in investigating, is a method for applying enrichments undertaken by DPLA as part of the ingestion process back to their local data sets. While DPLA provides data dumps for all Hubs' metadata both as individual and collective compressed dump files on the DPLA portal, working with this data can be challenging due—in part—to the sheer size of the files. For Hubs that have a strong technology team and a software environment that would allow it, pulling data from the DPLA API and merging changes with their local data might be a possibility. For others, especially those using systems like CONTENTdm that do not allow for the expression of relationships between fields, this will likely remain an impossibility. Nonetheless, to provide this service in a scalable fashion will require DPLA to better track how and when enrichments are applied, and when they may or may not be necessary.

4.4. Further tool and infrastructure development

While DPLA provides guidance to Hubs about particular standards, schemas, or protocols used to standardize, aggregate, and/or provide metadata, DPLA does not usually recommend or require use of any specific tools or applications to harvest, transform, or enrich metadata. Some Hubs have expressed an interest in working with other Hubs or with DPLA to develop tools to help with these processes. Even when formal collaboration has not yet been established, DPLA now finds itself providing an important service, mediating connections across Hubs to identify when the community faces common challenges.

5. Planning for needed improvements

Based on this feedback from Hubs, as well as needs identified through the challenges listed previously, DPLA is now reassessing its priorities and planning to address these issues. In some cases, resolving these issues may directly impact the infrastructure DPLA has in place, and addressing others clearly relates the need for DPLA to identify the level to which it should provide services on behalf of its Hubs. Some of the major areas of focused effort over the next year include the following.

5.1. Revision of the DPLA Metadata Application Profile

While the Metadata Application Profile is based on the Europeana Data Model (EDM), it has nonetheless diverged from it due to the pressures of DPLA's initial launch outlined above. Accordingly, DPLA is undertaking revision of the MAP to bring it back to closer alignment with EDM, which will allow the ingestion process to better associate URIs with given predicates in the MAP. As indicated in section 2.1, DPLA had sufficient needs that led to the development and implementation of MAP version 3.1. As an organization, DPLA has committed to reviewing the MAP on an ongoing basis, and is already planning for further changes to be included in MAP version 4. These include shifting to the class defined by EDM for spatial data (edm:Place), better support for controlled vocabularies for subject and genre statements, and investigating the addition of a class to provide support for annotation information. Future versions will also allow DPLA and other consumers of the ingested metadata to better incorporate annotations, either in the form of user-generated metadata, or automated output based on the results of transforms and enrichments during each ingest process.

5.2. Reassessment of "data quality" and "validation" in the context of DPLA

To provide better tools that ensure the validity and quality of metadata, there will need to be a clear understanding of how those terms are defined in the context of the DPLA/Hub collaboration. Lagoze et al. (2006) suggest that safe transforms are not necessarily scalable, and as such, DPLA and its Hubs must work together to clearly identify which remediation or augmentation processes add the most value to partners and other stakeholders. In addition, DPLA needs to determine whether validation against the MAP is a priority, and to have a clearer delineation of which party must provide the appropriate source data to fulfill the obligations of the MAP (i.e., DPLA, the Hub, or the partner). If explicit validation against the MAP becomes a priority for DPLA and its stakeholders, it will likely require the addition of a means to validate a set of statements against the constraints of the MAP as an RDF application profile. As a preliminary investigation, the co-authors have contributed use cases to the DCMI RDF Application Profiles Task Group.

5.3. Encouraging Hubs to undertake metadata transformation and enrichment locally and to develop appropriate tools

Since Hubs often know their metadata (and that of their partners) best, DPLA sees promise in Hubs taking on greater responsibility for metadata remediation, enrichment, and transformation to the MAP at the local level whenever possible. In many cases, DPLA has seen leadership in this area from Service Hubs, in particular (organizations or collaborative endeavors that aggregate metadata and provide services to several cultural heritage organizations, usually at a state or regional level). Some Service Hubs are already actively developing open source software to support these processes. Ultimately, software and infrastructure developed by the Hubs may benefit DPLA and its network further if it can be easily reused.

There are several notable examples of this leadership shown by Service Hubs. Developers at the Boston Public Library (2014) have developed a Ruby module for improved geocoding and reconciliation of geographic names against vocabularies, which is used to augment both their own data as well as data aggregated by Digital Commonwealth, the Service Hub for Massachusetts. University of Minnesota Libraries (2014a, 2014b, 2014c) are developing a suite of tools to harvest, transform, and augment metadata for materials aggregated by the Minnesota Digital Library, with the ultimate goal to provide DPLA with the metadata compliant with the MAP. In addition, the North Carolina Digital Heritage Center (NCDHC) has gained significant expertise in using REPOX for metadata aggregation as a DPLA Service Hub and has developed additional quality assurance applications to support this work (Gregory and Williams, 2014). In addition, to promote reuse, NCDHC released these as open source applications on GitHub. The tools allow NCDHC staff to review mappings, check for the presence of required properties or elements (NCDHC 2014a), and to provide a preview simulating the DPLA's portal user interface for individual new records that can be reviewed by their partners (NCDHC 2014b).

5.4. Improvement of documentation for metadata model and ingestion process

Despite both metadata mapping documentation and the code for the ingestion system being publicly available, there is still a significant gap in terms of materials available to understand the DPLA ingestion process. Accordingly, DPLA has begun to address this need by releasing an introductory white paper that explains the MAP (Digital Public Library of America 2014c) and creating a wiki page that collocates existing documentation about metadata, partnerships, and related activities (2014d). DPLA continues to develop further documentation that describes the ingestion process. This work will also likely give DPLA staff better insight about the expectations for these processes. In addition, DPLA staff has also supplemented the MAP version 3.1 documentation with explicit references to how properties within MAP are serialized as JSON-LD.

5.5. Improvement or replacement of the DPLA ingestion system

Many of the issues identified by DPLA demonstrate that the current ingestion system, while suitable as a prototype platform for the harvesting, remediation, mapping, and enhancement from many sources, is not entirely suited to the needs of a large-scale aggregator. Internally, DPLA staff has been working to address some issues while investigating whether a substantial refactor or a complete replacement would better serve the needs of the organization. A few areas for immediate focus include increasing efficiency, providing better automation, allowing DPLA content staff to oversee and understand the ingestion process directly with less mediation by the DPLA technology staff by the development of the aforementioned ingestion and QA dashboards, and more clearly defining the shared set of transforms and enrichments for all sources. In addition, the use of domain specific languages that are purpose-built for metadata mapping, transformation, and enhancement holds promise (e.g., Phillips, Tarver and Frakes, 2014 and LibreCat, 2014). These changes, in turn, could allow DPLA to create a system with its Hubs that is more approachable and transparent for those less comfortable with command-line applications and the orchestration of web services. DPLA has not committed to specific candidates for a replacement or undertaken extensive requirements analysis for a new ingestion system. Nonetheless, DPLA is interested in investigating both the previously described software suite under development by University of Minnesota, as well as Supplejack, the harvesting and augmentation framework used by DigitalNZ (2014).

6. Conclusion

Despite ongoing challenges with its existing infrastructure, DPLA has successfully aggregated over seven million records from 20 Hubs and nearly 1,300 partner institutions. The lightweight infrastructure used to support ingestion, storage, and indexing allowed the technical implementation team to quickly develop a system to harvest, remediate, and enrich metadata in varying formats. While the current ingestion system clearly has limits, the experience has allowed DPLA and its Hubs to identify shared needs and opportunities for collaboration while adding value to metadata for digitized cultural heritage materials. As the partnership around DPLA grows, the organization is uniquely situated to foster a community of practice that develops and provides documentation, software, and a forum to address ongoing needs in the remediation and enhancement of metadata at a national scale.

Acknowledgements

The Digital Public Library of America wishes to thank and acknowledge the support of the following organizations that have funded its efforts: The Alfred P. Sloan Foundation; The

Andrew W. Mellon Foundation; The Arcadia Fund; The Bill & Melinda Gates Foundation; the Institute of Museum and Library Services; The John S. and James L. Knight Foundation; The Mrs. Giles Whiting Foundation; and the National Endowment for the Humanities.

References

Akara. (2010). Retrieved August 7, 2014, from http://akara.info/.

DigitalNZ. (2014). Supplejack documentation, version 0.1. Retrieved August 7, 2014, from http://digitalnz.github.io/supplejack/.

Boston Public Library. (2014). Bplgeo. Retrieved August 7, 2014, from https://github.com/boston-library/Bplgeo.

Digital Public Library of America. (2014a). Digital Public Library of America Metadata Application Profile, Version 3.1. Retrieved August 7, 2014, from http://dp.la/about/map.

Digital Public Library of America. (2014b). The DPLA ingestion system, version 31.1. http://dx.doi.org/10.5281/zenodo.11226. Retrieved August 7, 2014, from https://github.com/dpla/ingestion.

Digital Public Library of America. (2014c). An introduction to the DPLA metadata model. Retrieved August 7, 2014, from http://dp.la/info/2014/03/25/intro-dpla-metadata-model/.

Digital Public Library of America (2014d). Content wiki. Retrieved August 7, 2014, from https://digitalpubliclibraryofamerica.atlassian.net/wiki/display/CT/Content.

DPLA RDF application profile use cases. (2014). Retrieved August 7, 2014, from http://wiki.dublincore.org/index.php/DPLA_RDF_application_profile_use_cases.

Europeana. (2013). Europeana Data Model primer. 14 July 2013. Retrieved August 7, 2014, from http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.

Europeana. (2014). Definition of the Europeana Data Model v5.2.5. 22 May 2014. Retrieved August 7, 2014, from http://pro.europeana.eu/documents/900548/0d0f6ec3-1905-4c4f-96c8-1d817c03123c.

Galiegue, Francis, Kris Zyp, and Gary Court. (2013). JSON Schema: interactive and non interactive validation. IETF Internet-Draft, January 30, 2013. Retrieved August 7, 2014 from http://json-schema.org/latest/json-schema-validation.html.

Gregory, Lisa, and Stephanie Williams. (2014). On being a hub: some details behind providing metadata for the Digital Public Library of America. *D-Lib Magazine, 20*(7/8). http://dx.doi.org/10.1045/july2014-gregory.

Hillmann, Diane I., Naomi Dushay, and Jon Phipps. (2004). Improving metadata quality: augmentation and recombination. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004. Retrieved May 15, 2014 from http://hdl.handle.net/1813/7897.

Lagoze, Carl, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom, and John Saylor. (2006). Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. In G. Marchionini, M. L. Nelson, and C. Marshall (Eds.): *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 230-239). New York: Association for Computing Machinery.

LibreCat. (2014). Catmandu: Introduction. Retrieved August 7, 2014, from https://github.com/LibreCat/Catmandu/wiki/Introduction.

NCDHC. (2014a). dpla-aggregation-tools. Retrieved August 7, 2014, from https://github.com/ncdhc/dpla-aggregation-tools.

NCDHC. (2014b). dpla-submission-precheck. Retrieved August 7, 2014, from https://github.com/ncdhc/dpla-submission-precheck.

Phillips, Mark, Hannah Tarver, and Stacy Frakes. (2014). Implementing a collaborative workflow for metadata analysis, quality improvement, and mapping. *Code4lib Journal, 23*. Retrieved August 7, 2014, from http://journal.code4lib.org/articles/9199.

Riley, Jenn, John Chapman, Sarah Shreeves, Laura Akerman, and William Landis. (2008). Promoting shareability: metadata activities of the DLF Aquifer initiative. *Journal of Library Metadata, 8*(3).

Sporny, Manu, Gregg Kellogg, and Markus Lanthaler (Eds.). (2014). JSON-LD 1.0: A JSON-Based Serialization of Linked Data. W3C Recommendation 16 January 2014. Retrieved August 7, 2014, from http://www.w3.org/TR/json-ld/.

University of Minnesota Libraries. (2014a). dpla.client. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.client.

University of Minnesota Libraries. (2014b). dpla.docs. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.docs.

University of Minnesota Libraries. (2014c). dpla.services. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.services.

Applying a Linked Data Compliant Model: The Usage of the Europeana Data Model by the Deutsche Digitale Bibliothek

Stefanie Rühle
Göttingen State and University Library, Germany
sruehle@sub.uni-goettingen.de

Francesca Schulze
German National Library, Germany
f.schulze@dnb.de

Michael Büchner
German National Library, Germany
m.buechner@dnb.de

Abstract

In 2013/14 the Deutsche Digitale Bibliothek (DDB) changed its data model from the CIDOC conceptual reference model to the Europeana Data Model (EDM). This decision was taken against the background of two major mandates the DDB has to fulfill: as a portal and as a platform the DDB is providing access to digital objects from German cultural heritage and research institutions. The DDB also aims to become the German aggregator for Europeana. Using EDM as the internal DDB data model was considered the most reasonable solution to meet these challenges. The DDB uses the data model for all portal functions that require semantic links between metadata (search facets, hierarchies, links between authority files and digital objects). The application of EDM for the DDB portal created some difficulties since not all necessary classes and properties had been entirely implemented in Europeana-EDM at that time. Therefore, DDB defined a metadata model which is based on the Europeana Data Model Definition but contains additional extensions. The DDB publishes metadata under the CC0 Public Domain Dedication license in EDM-RDF/XML via an OAI-PMH interface to serve Europeana and also via an Application Programming Interface (API) for external users to develop new applications on the basis of metadata harmonized by the DDB.

Keywords: Deutsche Digitale Bibliothek; German Digital Library; Europeana Data Model; CIDOC Conceptual Reference Model; metadata model; metadata mapping; metadata interoperability; linked data

1. Introduction

The Deutsche Digitale Bibliothek (DDB) provides a portal and a platform providing access to digital objects from German cultural heritage and research institutions. It brings together specialists from archives, museums, libraries as well as research, monument protection and media institutions in a Competence Network, funded by federal, state and local authorities. The full version of the portal was launched in March 2014. Besides being the main access point to digitized cultural and academic objects from Germany the DDB aims to become the German aggregator for Europeana, the central access point to Europe´s digitized cultural heritage. Europeana is operated by the Europeana Foundation and provides public services like the Europeana portal[1]. It accumulates and distributes metadata on digital collections from data providers across Europe, for example the DDB. Europeana encouraged the DDB to change the basis for its internal metadata model from CIDOC-CRM to the Europeana Data Model (EDM). EDM is a linked data compliant model developed by Europeana. It uses properties and classes of different namespaces, i. e. terms of the Dublin Core Metadata Element Set, the DCMI Terms and the OAI-ORE (EDM Definition, 2013). The DDB metadata model also uses properties and classes defined by Europeana taking into account the event-based modelling of object lifecycles

[1] URL to Europeana portal: http://www.europeana.eu/portal/

in CIDOC-CRM, however, these descriptions are less complex than in CRM (CRM Definition, 2014). In 2013/14 the DDB replaced CRM with EDM. As a result, mappings to the internal DDB format became less complex which reduces costs of metadata transformations. Using EDM also enables the reusability of Europeana tools. This report presents different applications on the basis of EDM in the DDB and describes the extensions of the model for DDB purposes. With this example, we want to illustrate that EDM is suitable as a domain model for the representation of digital cultural heritage. This model can also be used beyond the purpose of delivering metadata to Europeana. Other projects which adapted or extended EDM for their purpose are for instance The European Library[2], Digitised Manuscripts to Europeana[3] or Europeana Fashion[4].

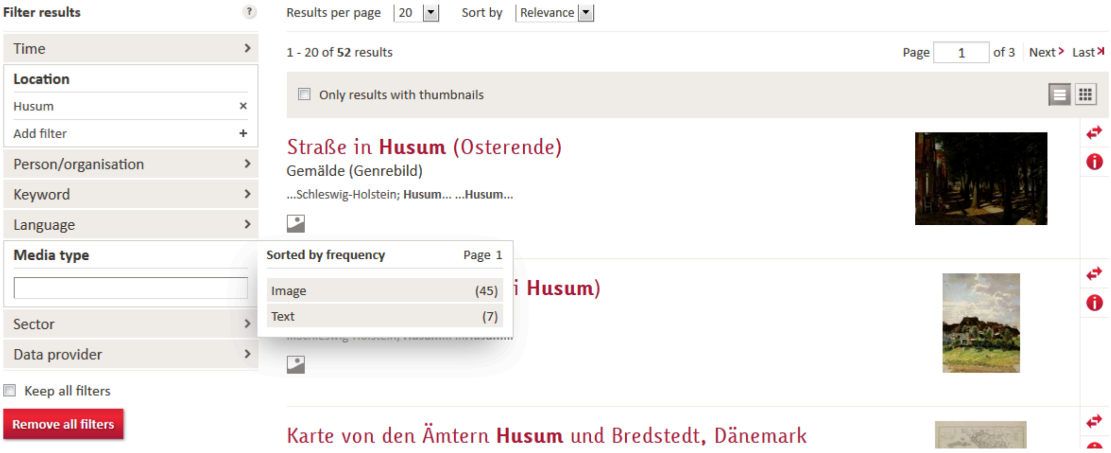
2. Use of EDM in the DDB

The requirements of the DDB concerning the data model are a result of the EDM triples' functions in the DDB. EDM in the DDB (in the following called DDB-EDM) is used

- for an advanced and facet-based search in the DDB portal,
- to represent the hierarchical organization of the digitized objects,
- to interlink objects and authorities, and
- to publish the data via OAI-PMH and an Application Programming Interface (API).

2.1. Facets

The facet-based search enables users to filter their search results by means of defined categories.



Results per page 20 Sort by Relevance

1 - 20 of 52 results Page 1 of 3 Next > Last

Straße in Husum (Osterende)
Gemälde (Genrebild)
...Schleswig-Holstein; Husum... ...Husum...

i Husum)

Karte von den Ämtern Husum und Bredstedt, Dänemark

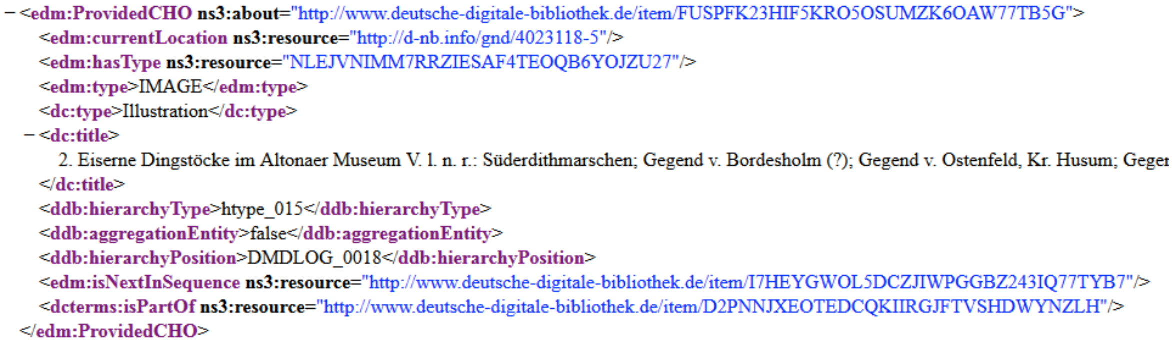
FIG. 1. Facets in the DDB Portal

The categories are based on the classes edm:TimeSpan for time, edm:Place for location, edm:Agent and dcterms:ProvenanceStatement for person/organization and data provider, skos:Concept for keyword, media type and sector, and dcterms:LinguisticSystem for language. In a next step, some of these categories will be refined using triples and controlled terms specifying the relation between an object and a place, time, person or organization.[5] This will allow users to distinguish between the "aboutness" of an object and information concerning its lifecycle and help them to differentiate whether it is the time and place of creation or modification, whether a person was involved in the finding or the destruction of an object etc.

[2] For a project description, see http://dm2e.eu.
[3] For a project description, see http://www.theeuropeanlibrary.org/tel4/.
[4] For a project description, see http://www.europeanafashion.eu/portal/home.html.
[5] For the specification of these relations the DDB uses URIs of the event vocabulary developed by the LIDO Community, see http://terminology.lido-schema.org/eventType.

2.2. Hierarchy

To describe the hierarchical relations between objects – e. g. the hierarchy of resources from libraries or archives – we use two classes to express the different nodes of a hierarchy in the user interface: the edm:ProvidedCHO for objects with a proper name or description (e. g. monographs, journals, articles, illustration) and the edm:PhyscialThing for nodes that are not described with a proper name or description but are needed to express the hierarchical structure (e. g. an issue). We use a domain specific property called `ddb:hierarchyPosition` for the description of the order of resources inside the hierarchy. Besides this property, DDB-EDM includes `edm:isNextInSequence` for compliance with the EDM used in Europeana.[6]



```
– <edm:ProvidedCHO ns3:about="http://www.deutsche-digitale-bibliothek.de/item/FUSPFK23HIF5KRO5OSUMZK6OAW77TB5G">
    <edm:currentLocation ns3:resource="http://d-nb.info/gnd/4023118-5"/>
    <edm:hasType ns3:resource="NLEJVNIMM7RRZIESAF4TEOQB6YOJZU27"/>
    <edm:type>IMAGE</edm:type>
    <dc:type>Illustration</dc:type>
  – <dc:title>
      2. Eiserne Dingstöcke im Altonaer Museum V. l. n. r.: Süderdithmarschen; Gegend v. Bordesholm (?); Gegend v. Ostenfeld, Kr. Husum; Geger
    </dc:title>
    <ddb:hierarchyType>htype_015</ddb:hierarchyType>
    <ddb:aggregationEntity>false</ddb:aggregationEntity>
    <ddb:hierarchyPosition>DMDLOG_0018</ddb:hierarchyPosition>
    <edm:isNextInSequence ns3:resource="http://www.deutsche-digitale-bibliothek.de/item/I7HEYGWOL5DCZJIWPGGBZ243IQ77TYB7"/>
    <dcterms:isPartOf ns3:resource="http://www.deutsche-digitale-bibliothek.de/item/D2PNNJXEOTEDCQKIIRGJFTVSHDWYNZLH"/>
  </edm:ProvidedCHO>
```

FIG. 2. Description of an edm:ProvidedCHO in DDB-EDM

2.3. Interlinking with Authorities

We use EDM to interlink DDB objects with resources from external data sources. As a first step, we connected DDB objects with person authority files from the Integrated Authority File (Gemeinsame Normdatei, GND). To establish the relations we exploit only GND URIs which are delivered in the original metadata. For persons, who play a role in the lifecycle of an object (e. g. author), we extended EDM with the CIDOC-CRM-Property `P11_had_participant`. For the inverse relation, i. e. from a GND person to DDB objects, we use the EDM property `edm:wasPresentAt`. Furthermore, we use the Dublin Core property `dcterms:subject` for persons, who are described or depicted by the object. To exploit information behind respective GND URIs and to offer person pages in the DDB portal, we apply the web service Entity Facts[7] offered by the German National Library. It allows other applications to integrate and interlink information from GND entities with their data sources. Entity Facts is implementing data enrichment, therefore different data sources (e. g. external links from BEACON files or images of persons from Wikipedia) are merged into a simple and easy-to-use JSON-LD fact sheet. The first version delivers information on entities of the GND entity type Person via an API. Subsequent versions will supply information on places and corporations as well. The GND is widely used in the library community and less represented in other sectors. Therefore, the DDB is developing an assessment tool[8] that will support users to compare, match and map their domain-specific vocabularies to the GND in a semi-automatic way.

2.4. Publication as Linked Data

We provide metadata of the cultural heritage institutions in the DDB-EDM RDF/XML format by applying linked data principles. We use URIs to uniquely identify different resources and their relations in RDF. Therefore, we transfer URIs from the original metadata records during the mapping to EDM whenever possible. Apart from the GND, we take URIs from vocabularies

[6] For information about hierarchies in Europeana see Task Force on hierarchical objects, 2013.
[7] For an example see the query for "Johann Wolfgang von Goethe" at http://hub.culturegraph.org/entityfacts/v1/118540238.
[8] The assessement tool is developed by digiCULT, a project partner of the DDB, see http://www.digicult-verbund.de/.

which are available as Linked Open Data, like Iconclass[9], Dewey Decimal Classification[10] or the Library of Congress vocabularies[11]. We also create URIs, for instance by adding a namespace to a code or identifier provided in the original metadata record (e.g. ISO 639-2 code "eng" to "http://id.loc.gov/vocabulary/iso639-2/eng"). Moreover, we include URIs from the ddb-vocnet namespace into EDM properties to receive controlled terms for the search in the DDB portal. This affects mostly properties, which express the type of a resource (e.g. type of a digital representation of an object). For the identification of some resources, however, it was necessary to additionally establish DDB-internal URIs. These URIs have a DDB-namespace and are created on the basis of common rules for respective DDB resources (e. g. resource class name/ISIL[12]/local identifier). In order for external users to recognize non-resolvable DDB-internal URIs they are encoded by a hash (e.g. EO5NPTOTBJL4V3RXVRLXE7YME7HY6DCW as can be seen in figure 2).

DDB-EDM RDF/XML records contain the results of our normalization and enrichment processes. An example is the use of DDB license URIs for both the metadata record and the digital object. The DDB licenses, which are compliant with the Europeana Licensing Model, give external users information whether and how they can reuse the metadata and digital objects. The DDB publishes its metadata records under the CC0 Public Domain Dedication license via its API[13]. This allows the development of further applications by using DDB metadata. Even though the DDB-API supplies the metadata in different XML formats (source format, DDB-EDM, DDB-View), DDB-EDM is considered as the most harmonized, interlinked and enriched representation of the metadata describing the objects. An application on the basis of the DDB-API is "Archivportal-D[14]" – a portal which provides a view on the DDB content and metadata from an archival perspective. DDB also delivers EDM metadata sets under the license CC0 via an OAI-PMH interface to Europeana. The interface is open to the public as well.

3. Mapping Workflow

The workflow to integrate metadata sets from institutions into the DDB consists of three main steps: 1) clarification of formal and content-specific aspects, 2) data clearing, and 3) ingest. An institution willing to participate has to fill out a content questionnaire including information about the holding/collection and the metadata format (MARC, METS/MODS, ESE, EAD, LIDO et al.). The data clearing begins with the analysis of test data and the adjustment of mapping rules. The original metadata is transformed with XSLT scripts to all DDB target formats, comprising EDM. All metadata representations of an object record are structured in the container format Cortex defined by the DDB. After the ingestion into the DDB test system, data experts review the quality of the transformation result in the test portal and in an XML preview. To support quality control, the DDB is implementing a validation tool. Data clearing is an iterative process with several circles of reviews and adjustments. After approval by the data provider the complete data contribution is ingested into the DDB backend and published via the DDB frontend (portal) and other public interfaces.

The switch from CRM to EDM had a strong impact on our mapping workflow and back-end operations. Since the data sets from all providers that were published via the DDB at that time had to be represented in the new DDB-EDM data format we had a big one-time effort to adjust all respective steps in our workflow. These were: a) the definition of new rules to map the elements and their contents from seven source formats to EDM, b) the indication of provider specific

[9] A classification system for art and iconography. For further information see http://www.iconclass.nl/home.

[10] See http://dewey.info/.

[11] See http://id.loc.gov/.

[12] ISIL is an acronym for International Standard Identifier for Libraries and Related Organisations. The registration for German institutions is managed by the German ISIL and Library Codes Agency at the Staatsbibliothek zu Berlin.

[13] The API of the DDB is documented in the wiki space "API der Deutschen Digitalen Bibliothek", available under the URL: https://api.deutsche-digitale-bibliothek.de/doku/display/ADD/API+der+Deutschen+Digitalen+Bibliothek.

[14] The development of Archivportal-D is funded by Deutsche Forschungsgemeinschaft (DFG). The portal will be launched publicly in September 2014. For a project description in German language see http://www.landesarchiv-bw.de/web/54267.

information in the mappings, c) the adaptation of the transformation tools including the programming of new XSLT scripts, d) the adjustment of the SOLR schema, e) the configuration of the search facets and hierarchies for the frontend, f) the transformation, ingestion and indexing of the complete DDB holdings which comprised around six million records in 2013.

Even though we installed a process that ensured that CRM and EDM records could be ingested in parallel, a few concessions had to be made. For instance, we prioritized the change of the published data sets to EDM. This resulted in a slower increase of content in the DDB since little resources were left for new ingests.

However, the introduction of DDB-EDM decreased the workload for the conceptual and technical mappings considerably. The establishment of mappings to CRM required expert knowledge. Our domain experts, however, were more familiar with EDM because they were already involved in mapping activities for contributing metadata to Europeana via other projects. Furthermore, with EDM the mappings became less complex and less error-prone, because in CRM a statement can be expressed in many ways which often resulted into a series of triples. For example, to state that an object is about a person the mapper had to opt for one of the following paths in CRM:

- E89 Propositional Object (or Subclass) P67F refers to E39 Actor (or Subclass)
- E89 Propositional Object (or Subclass) P129F is about E39 Actor (or Subclass)
- E24 Physical Man-Made Thing (or Subclass) P62F depicts E39 Actor (or Subclass)

We map this statement to DDB-EDM as follows:

- edm:ProvidedCHO dcterms:subject edm:Agent

This example shows that we lost precision in DDB-EDM regarding semantic relations, because the CRM properties "refers to", "is about" and "depicts" were merged into the single EDM property "dcterms:subject". But this generic property is sufficient to distinguish the "aboutness" from the lifecycle of an object which is the crucial requirement for our search facets. This decision was also reasonable regarding the time saved for mappings, the processing of records and thus the ingestion of data contributions.

4. The DDB-EDM Model

The decision to minimize the transformation costs by using EDM in the DDB raised some difficulties. Coming from the event based CIDOC-CRM, the DDB needed properties and classes to describe the events in the lifecycle of the digitized resource. Such properties and classes were available in EDM, but at that time Europeana had not yet implemented them entirely, especially not the necessary event class and its associated properties. Therefore we developed a DDB-EDM model that was an extension of the implemented Europeana EDM described in the Europeana Mapping Guidelines (EDM Mapping Guidelines, 2013).

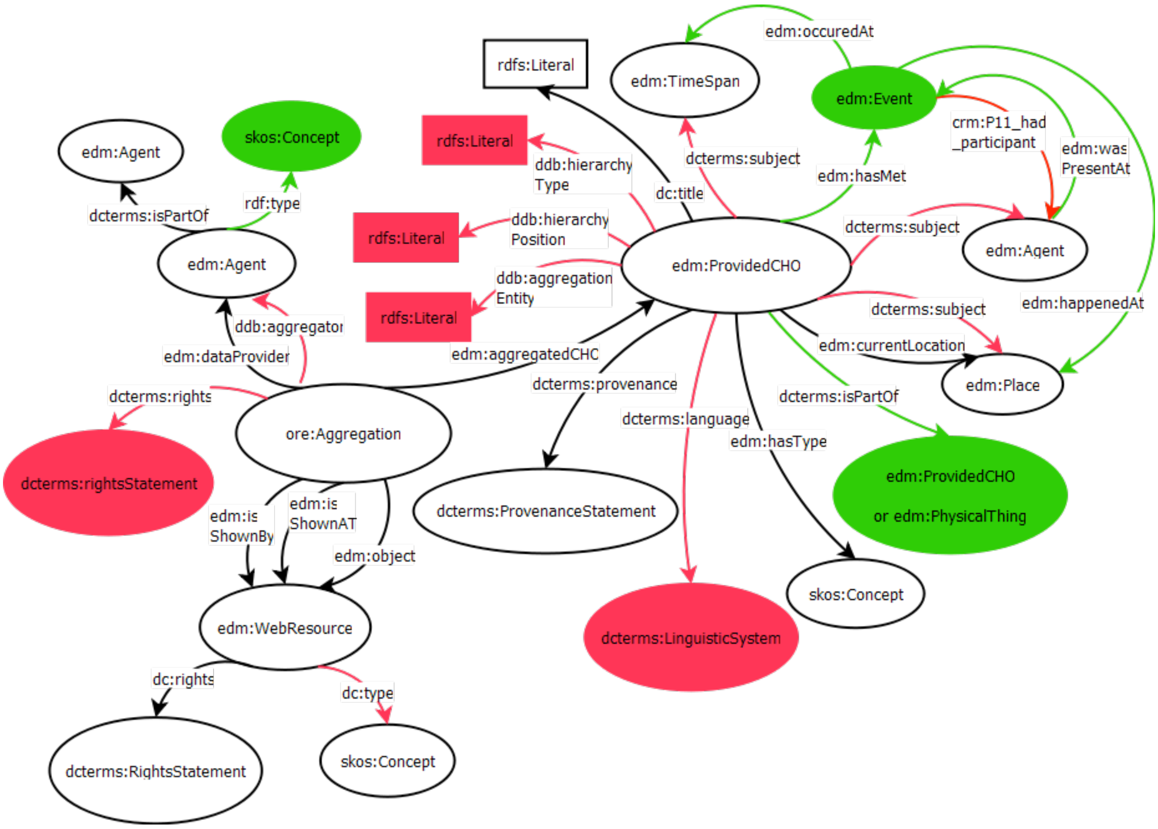


FIG. 3. DDB-EDM model

Figure 3 gives an overview over the properties and classes used in the DDB. Properties and classes used in the DDB, but not implemented in Europeana in 2013, are colored green. This concerns all statements about `edm:Event` and `edm:PhysicalThing`. Properties and classes used in the DDB, but not compliant with the Europeana Model, are red. We use these terms for domain specific requirements. These are:

- `dcterms:rights` with `ore:Aggregation` as domain and `dcterms:RightsStatement` as range, used for rights statements about the metadata. Depending on the value of this property the metadata will be provided by the DDB to Europeana or not[15],
- `ddb:aggregator` with `ore:Aggregation` as domain and `edm:Agent` as range, used for the aggregator providing data to the DDB[16],
- `dcterms:rights` with `edm:WebResource` as domain and `dcterms:RightsStatement` as range, used for DDB specific rights statements,
- `dcterms:language` with `edm:ProvidedCHO` as domain and `dcterms:LinguisticSystem` as range, used to describe the language of the resource with non-literal values[17],
- `dcterms:subject` with `edm:ProvidedCHO` as domain and a non-literal value as range which may be an instance of one of the EDM conceptual classes `edm:Agent`, `edm:Place`, `edm:TimeSpan` etc.[18],
- `ddb:hierarchyType` with `edm:ProvidedCHO` as domain and a literal value as range, used to describe the object type of an `edm:ProvidedCHO` or `edm:PhysicalThing` as part of a

[15] Metadata are only exposed to Europeana or others when the value is "CC0".

[16] Europeana uses edm:provider for Europeana aggregators which in our case is the DDB. Because the property is not repeatable the DDB needs a domain specific property for the description of DDB aggregators.

[17] Europeana uses dc:language and allows the use of literal and non-literal values whereas the use of URIs in the DDB is mandatory.

[18] Europeana uses dc:subject and allows the use of literal and non-literal values whereas the use of URIs in the DDB is mandatory.

hierarchy (e.g. journal, volume, article, illustration). Values used here are based on a vocabulary that will be published as Linked Open Data in the future, which will result in a revision of the DDB-EDM model,

- `ddb:hierarchyPosition` with `edm:ProvidedCHO` as domain and a literal value as range, used to describe the order of an `edm:ProvidedCHO` or `edm:PhysicalThing` in a hierarchy,
- `ddb:aggregationEntity` with `edm:ProvidedCHO` as domain and a literal value as range, used to distinguish between hierarchical levels with proper descriptions and levels without such descriptions (e.g. an issue that is only identified by the number),
- `rdf:type` with `edm:Agent` as domain and `skos:Concept` as range, used to describe the relation between a corporate body and the type of sector it belongs to, and
- `crm:P11_had_participant` with `edm:Event` as domain and `edm:Agent` as range, used to describe that there is a relation between an event and an agent (e. g. the creation event and the creator).

5. Conclusion and Outline

The implementation of EDM has turned out to be the most effective way to serve the requirements of the DDB portal for functions based on linked data principles and external applications like Europeana. Prospectively, DDB-EDM will also contain the results of further enrichment and normalization processes the DDB is currently establishing for authority data and controlled vocabularies which will subsequently improve the portal as well.

References

CRM Definition (2014). Definition of the CIDOC Conceptual Reference Model, Version 5.1.2. Retrieved April 24, 2014 from http://cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf

EDM Definition (2013). Definition of the Europeana Data Model, version 5.2.4. Retrieved April 24, 2014 http://pro.europeana.eu/edm-documentation

EDM Mapping Guidelines (2013). Europeana Data Model – Mapping Guidelines, Version 2.0. Retrieved April 29, 2014 from http://pro.europeana.eu/edm-documentation

Task Force on hierarchical objects (2013). Recommendations for the representation of hierarchical objects in Europeana. Retrieved April 29, 2014 from http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Taskforce+on+hierarchical+objects

Distributed Metadata Environments & Aggregation— Part B

Designing a Multi-level Metadata Standard based on Dublin Core for Museum data

Jing Wan
Beijing University of Chemical Technology, China
wanj@mail.buct.edu.cn

Yubin Zhou
Beijing University of Chemical Technology, China
2011200868@grad.buct.edu.cn

Gang Chen
Beijing Gehua Culture Development Group, China
cg@sach.gov.cn

Junkai Yi
Beijing University of Chemical Technology, China
yijk@mail.buct.edu.cn

Abstract

Metadata is a critical aspect of describing, managing and sharing museum data. It is challenging to develop a general metadata schema that meets the requirements of different museums due to the large range of data types. The capability of concise description and the simplicity of use need to be considered. In this paper, we report on a finished project that aims to design a metadata schema for museums in China. An extensible metadata standard based on Dublin Core is presented, which includes core of metadata, extension rules and specific metadata. For the core metadata, we introduce terms, definitions, registration rules and detailed examples of description. The principle of choosing terms and refinements is discussed. A specific metadata schema for porcelain is discussed as an extension example.

Keywords: metadata; Dublin Core; museum

1. Introduction and Motivation

With the rapid development of information technology since the 1992, lots of museums adopted collection management systems, digitalized collection data, and provided public. Data sharing and integration among museums became important.

Metadata is defined as "structured data about data". As a key issue of data standardization and data sharing, metadata for cultural heritage has attracted worldwide attention. A number of organizations and initiatives made great efforts to address this issue. Some published metadata schemas have been widely used and accepted as international standards, for example, Dublin Core (DCMI, 2012), CDWA (Getty Research Institute, 2008), EDM (Europeana Foundation, 2013), CIDOC CRM (CIDOC CRM Special Interest Group, 2011), VRA Core (Visual Resources Association Data Standards Committee, 2007), EAD (Society of American Archivists and the Library of Congress, 2002), and FGDC/CSDGM (Federal Geographic Data Committee, 1988).

China's management system of cultural relics is different from that of other countries. Most cultural relics are owned by the state and under the protection of the state. A state department takes charge of the work concerning cultural relics throughout the country. From 1978, a serial of regulations were published by China's State Administration of Cultural Heritage, which aimed to establish the standard process in registering and compiling files for museum collections. Many government funded projects promoted the work of museum informatics. The project "Cultural Relics Census and Collection Management System Construction" started in 2001, with 48,006 pieces of valuable collections and 1,370,000 pieces of general collections recorded in the database by 2010. In 2012, the project "First National Movable Cultural Relics Census" started, which aimed to investigate, identify and register movable relics through information technology.

Many museums in China established collection management systems and digitized their collections progressively, such as, the Palace Museum, the Capital Museum, and Shanghai Museum. Some museums designed their own data specifications. And several specifications were published by the government, for example, "Data Specification for Museum Collections", "Standard for Image Archive of Unmovable Cultural Relics", "Data Specification for the Third National Heritage Sites Census", and "Data specification for the First National Movable Cultural Relics Census".

But there is still no national standard for museum data in China. Considering the different management systems, it is difficult to utilize existing metadata schemas without modification. And museums are different in collection types, collection quantities, data quality and the skill levels of staff. So different requirements for metadata need to be considered. The metadata schema should be capable of concise description, be simple to use, and be compatible with the published specifications.

We describe an effort in developing a metadata architecture to address this issue. In the project, we design the core metadata based on Dublin Core, and specific metadata extensions for drawings, porcelain, ancient buildings and inscriptions. For each metadata of these categories, we provide terms, definitions, refinements, registration rules and detailed samples. In this paper, we focus on the core metadata and describe one specific example of metadata extension.

2. Metadata Architecture

Figure 1 shows the metadata architecture, which includes the core metadata, specific metadata and extension rules.

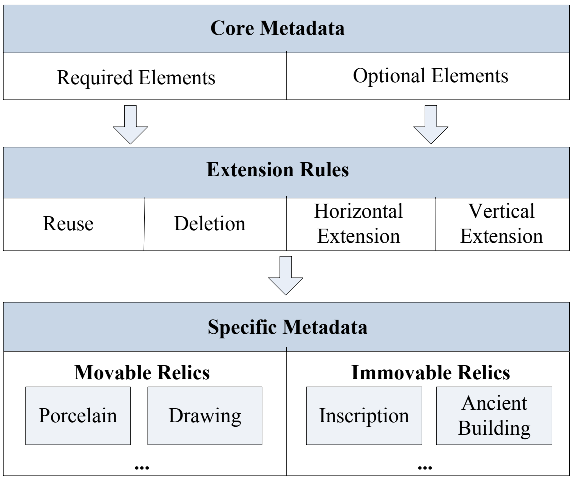


FIG. 1. Metadata architecture

The core metadata is simple and based on Dublin Core Metadata Element Set, version 1.1 (Dublin Core, 2012). This level is used to describe the general core attributes of digital resources. It supports retrieval, integration and data exchange. The elements of this level are easy to use. A museum that has simple data could use it directly. And a museum with a large number of collections and complex data structures can use it as the first stage of a plan. For these museums, entering complete data usually takes several years or even decades. First make it work, and then make it better. This rule is helpful to motivate the staff, get support from other divisions, and gain experience.

The specific metadata is used for data sets of particular type or domain. It is designed by analyzing existing archives and possible data requirements coming from museum management.

The extension rules are used to extend metadata to meet the actual requirements of a specific museum. Rules and implement approaches need to be provided to guide users in customizing metadata.

3. Core Metadata

3.1. Approach

The Core Metadata consists of an elements set and qualifiers. It is a vocabulary of nineteen properties for use in digital museum's collection description. "Core" means its elements are generic, and usable for describing a wide range of museum data.

Taking into account versatility, scalability, and interoperability, we design the core based on the Dublin Core Metadata Element set. In addition, data specifications and published standards in China are considered. Existing data are stored in the database or on paper are analyzed. We also consider the data elements adopted by the cultural relics census. Using this approach, we adopt eleven elements from Dublin Core and add eight elements and qualifiers.

Element qualifiers make the meaning of an element narrower or more specific. Following the practice of Dublin Core Qualifiers, there are two classes of qualifiers, element refinements and encoding schemes. The element refinements include object qualifier, basic qualifiers, and composite qualifiers.

1. ***Object qualifier***. The metadata should be capable of describing movable and immovable cultural relics. But these two types of relics have great differences. This qualifier is used to describe the range of an element.
2. ***Basic qualifier***. It is the basic unit of qualifier. It cannot be extended.
3. ***Composite qualifier***. It consists of basic qualifiers and/or composite qualifiers. For example, the copyright of the image has a composite qualifier including three basic qualifiers—owner, copyright restriction, and copyright description.

We define each element and qualifier by nine properties, which are name, identifier, version, definition, repeatability, data type, required status, domain, and qualifier.

3.2. Element Set

The element set of the core metadata includes nineteen terms. We adopt terms from Dublin Core Metadata Element Set, version 1.1 with the exception of *language*, *contributor*, *publisher* and *source* (DCMI, 2012). For collections in China, the *language* element always has the value "in chinese", so we don't adopt it now. The *contributor* element and the *publisher* element of a collection are the same as its keeper, which is included in the element *rights*. So we don't adopt *contributor* or *publisher*. We ignore the element *source* for it has no value for a collection. Table 1 shows the correspondence between the core metadata elements and the Dublin Core metadata elements. Many of these terms have basic constraints.

We describe standard vocabularies for some elements. The value "Yes" of the Encoding Scheme column of Table 1 on the following page indicates that vocabularies for the element are provided. For example, the *grade* of the movable cultural relics includes the values "grade one", "grade two", "grade three", "not determined", and "normal". These terms are defined in the standard "Grading Standard For Cultural Relics" published by China's Ministry of Culture.

TABLE 1: Alignment of the core metadata element set and DC element set.

Term	Comment	Refinements	Encoding Scheme
Name	DC: Title	Registered Name, Alternative Name	
Identifier	DC: Identifier		
Type	DC: Type		Yes
Date	DC: Date		Yes
Subject	DC: Subject		
Description	DC: Description		
Creator	DC: Creator		
Coverage	DC: Coverage	Geographic Coordinate, Scope Coordinates(Measure point number, Measure Point Coordinates, Adjacent Measure point), Geographic Name	
Right	DC: Rights	Ownership Type, Affiliation	Yes
Relation	DC: Relation	Image, Reference, Component	
Material	DC: Format	Material Type, Specific Material	Yes
Acquisition		Approach, Enter Scope, Enter Date	Yes
Grade			Yes
Measurement		Dimension(Length, Width, Height), Weight, Distribution Area, Protection Scope Area, Building Area, Construction Control Zone Area	
Conservation		Residual Level, Conservation Status, Status Assessment,	
Quantity			
Condition		Use Unit, Subordination Unit	Yes
Environment		Natural Environment, Humanities Environment	
DamageCause		Natural Cause, Man-made Cause	Yes

4. Extension Rules

Because of the large range of museum collections, it is hard to use the core metadata to meet the description of each item. So we design the extension rules to generate more specific metadata. And we provide the design of four specific metadata, which includes terms, definitions, registration rules and detailed examples.

There are four classes of extension approach:

Reuse. It refers to adopting existing elements or refinements of the core metadata. It includes complete reuse and partial reuse. The reuse class indicates adoption without modification. Partial reuse adds some restrictions.

Deletion. Refers to deleting elements or refinements that are useless in this level.

Horizontal extension. Refers to adding a new element.

Vertical extension. Refers to adding refinements according to the extension rules.

5. Metadata for Porcelain

The specific metadata for porcelain is an example of how the extension rules are applied. Table 2 shows how the specific metadata for porcelain is extended from the core metadata. It includes sixteen elements. The followings are examples of four extension rules with the porcelain metadata:

1. ***Reuse***. The element name and its two refinements (registered name and alternative name) from the core metadata are included in the specific metadata. It is complete reuse. The element grade from the core metadata is included in it too. But the value range of the element grade is changed , so it is part reuse.
2. ***Deletion***. The element coverage has three refinements in the core metadata. We delete one refinement (scope coordinates) in the specific metadata for it is useless for porcelain.

3. ***Horizontal extension.*** There is no horizontal extension.
4. ***Vertical extension.*** The element name has a new refinement (original name). We add it because many original names of porcelain collections are revised in order to conform to the naming rules published by the authority. The revised name is the registered name of a collection. But sometimes the original name is well known. So we need to record it too.

TABLE 2: Specific metadata for porcelain.

Index	Term	Refinements	Extension
1	Name	Registered Name, Alternative Name, Original Name	Complete Reuse+vertical Extension
2	Identifier		Complete Reuse
3	Type		Part Reuse
4	Date	Manufacture Date, Use Date	Vertical Extension
5	Subject		Vertical Extension
6	Description	This term has 17 refinements.	Vertical Extension
7	Creator	Name, Gender, Native Place, Birth, Death, Creator Description	Vertical Extension
8	Coverage	Geographic Coordinate, Geographic Name	Deletion+Complete Reuse
9	Right	Ownership Type, Affiliation	Complete Reuse
10	Relation	Image, Reference, Component	Complete Reuse+Vertical Extension
11	Material	Material Type, Specific Material	Complete Reuse
12	Acquisition	This term has 12 refinements.	Complete Reuse+Vertical Extension
13	Grade		Part Reuse
14	Measurement	Dimension (Length, Width, Height), Weight	Deletion+Complete Reuse
15	Conservation	Current Condition, Natural Damage, Physical Damage, Remarks, Citations	Complete Reuse+Vertical Extension+Deletion
16	Quantity		Complete Reuse

5. Conclusions and Future Work

This paper introduces a project aimed to design an extensible metadata standard for museum data in China. We consider the capability of concise description and the simplicity of use. We present a standard including core metadata, extension rules, and specific metadata. The core metadata is based on Dublin Core and is easy to use. It includes nineteen elements and refinements. There are four extension approaches that are reuse, deletion, horizontal extension and vertical extension.

In the future, we plan to develop a metadata management system, which will help museums to customize the metadata element set for their application. We also plan to enhance the use of standard vocabularies and make them compatible with the international standards.

Acknowledgements

We would like to express our thanks to the Capital Museum for providing data samples. We also gratefully appreciate the contribution of Hu Chui from the Palace Museum, Xiao Fei from the National Museum of China, Shen Guihua from China Cultural Heritage Information and Consulting Center for the discussion on the metadata. This work is financed by the State Cultural Relics Bureau within the project "*A Study of a Metadata Standard for Museum Data in China*" .

References

China's State Administration of Cultural Heritage. (2001). The Data Specification for Museum Collections. Retrieved April 1, 2014, http://www.nach.gov.cn/art/2008/7/9/art_343_3636.html.

China's State Administration of Cultural Heritage. (2005). The Standard for Image Archive of Unmovable Cultural Relics. Retrieved April 1, 2014, http://www.sach.gov.cn/art/2008/7/8/ art_343_3633.html.

China's State Administration of Cultural Heritage. (2007). The Data Specification for the Third National Heritage Sites Census. Retrieved April 1, 2014, from http://pucha.sach.gov.cn/tabid/64/Default.aspx.

China's State Administration of Cultural Heritage. (2012). The Data Specification for the First National Movable Cultural Relics Census. Retrieved April 1, 2014, http://www.wenwu.gov.cn/kydwwpc/.

CIDOC CRM Special Interest Group (SIG). (2011). Definition of the CIDOC CRM. Retrieved April 1, 2014, from http://www.cidoc-crm.org/definition_cidoc.html.

DCMI. (2012). DCMI Metadata Terms. Retrieved April 1, 2014, from http://dublincore.org/documents/dces/.

Europeana Foundation. (2013). EDM Definition. Retrieved April 1, 2014, from ehttp://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.

Federal Geographic Data Committee. (1998). Content Standard for Digital Geospatial Metadata. Retrieved April1, 2014, from http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html.

Getty Research Institute. (2008). Categories for the Description of Works of Art (CDWA). Retrieved April 1, 2014, http://www.getty.edu/research/publications/electronic_publications/cdwa/.

Metadata Architecture and Application Team and Taipei National Palace Museum.(2007). The Information Requirements Specification for Taipei National Palace Museum. Retrieved April 1, 2014, from http://www2.ndap.org.tw/eBook08/showContent.php?PK=51.

Society of American Archivists and the Library of Congress, 2002). Encoded Archival Description. Retrieved April 1, 2014, http://www.loc.gov/ead/.

Visual Resources Association Data Standards Committee. (2007). VRA Core 4.0 Introduction. Retrieved April 1, 2014, from http://www.loc.gov/standards/vracore/schemas.html.

"Lo-Fi to Hi-Fi": A New Way of Conceptualizing Metadata in Underserved Areas with the eGranary Digital Library

Deborah Maron
UNC Chapel Hill, USA
maron@live.unc.edu

Cliff Missen
UNC Chapel Hill, USA
missenc@unc.edu

Jane Greenberg
Metadata Research Center
janeg@email.unc.edu

Abstract

Digital information can bridge age-old gaps in access to information in traditionally underserved areas of the world. However, for those unfamiliar with abundant e-resources, their early exposure to the digital world can be like "drinking from a fire hose." For these audiences, abundant metadata and findability, along with easy-to-use interfaces, are key to their early success and adoption. To hasten the creation of metadata and user interfaces, the authors are experimenting with "crowd cataloging." This report documents their experimental and intended work and Maron's Lo-Fi to Hi-Fi metadata pyramid model guiding a developing metadata initiative being pursued with the eGranary Digital Library, the technology used by Widernet in a global effort to ameliorate information poverty. The work in development, the Lo-Fi to Hi-Fi model, has principles adapted from technical design processes and tried and true methods within metadata creation such as crowdsourcing. It attempts to reconceptualize the metadata modeling paradigm and aligns with research that has shown that community-based librarians are better poised to identify culturally congruent resources, but many require significant training in metadata concepts and skills. The model has amateurs (mostly students) crowdsource "lo-fi" terms, which domain experts and information professionals can curate and cull in "hi-fi" to enhance findability of resources within the eGranary while simultaneously honing their own computer, information and metadata literacies. Though the focus here is on Africa, the findings and practices can be universalized to off-line collections around the globe.

Keywords: information literacy; computer literacy; WiderNet; eGranary; Africa; metadata literacy; crowdsource; crowd catalog; folksonomy; LIS education; industrial design; hi-fi prototype; lo-fi prototype

1. Introduction

Many citizens of first world nations have become accustomed to, and even routinely take for granted, accessing information immediately and easily through the Web and mobile technologies. With years of experience under their belts, they blithely operate search engines, barely recalling how they developed their search skills over hundreds of hours of Internet use. However, not all individuals have that luxury, with over five billion people lacking access to Internet resources. Information poverty, begotten of a lack of technology and knowledge of how to use it, is a pervasive problem that affects quality of life, as well as development of crucial 21st century skills like information and computer literacy, for children and adults around the globe.

Lack of the aforementioned literacies is an information and library science (ILS) issue that impacts many librarians and library workers in all aspects of work including metadata creation and management. More specifically "metadata literacy" is adversely affected by the conditions in environments in which workers lack sufficient computer or information literacy.

The focus of the work reported here is the rural, indigenous sub-Saharan library, where a dearth in literacy is exacerbated by a lack of connectivity to the Internet. Driving questions include the following: How should metadata for essential resources be developed in these and similarly afflicted regional libraries? Also, how can the knowledge and terminology of a particular society be leveraged within these regions to create metadata? We propose that concepts from the informal technical design process employed by WiderNet over the last five years to develop topical portals for end users (in medicine, nursing, public health, rural agriculture, life skills, disability rights, etc.) be formally expanded and evaluated to provide a first-rate framework for conceptualizing and delivering metadata formation for people in underserved communities. This prototyping was further defined as "hi-fi" and "lo-fi" and extrapolated into a model by Maron in 2014.

We present this model in the context of WiderNet@UNC, an initiative to bring Internet content to places worldwide that are not connected, via massive hard disks of material that mirrors what the Internet has available. WiderNet's hard disks of material, called eGranary Digital Libraries, are currently used in over 800 locations and contain 35 million items each; while the majority of the documents can be searched using a built-in term search engine, only a fraction of the items are catalogued because the onus falls on the small team of paid and volunteer cataloguers to create records.

In evaluating logs from dozens of eGranary servers, it has been noted that users, generally unfamiliar with search engines, are more likely to use the limited catalog to locate resources. In most cases, 90-95% of the documents retrieved were listed in the catalog. Clearly, new users prefer well-cataloged resources.

This project report details the expanded concept of the Lo-Fi to Hi-Fi metadata pyramid, representing a method by which metadata can be crowdsourced and curated for resources by the very people that use and operate eGranaries within underserved areas of Africa in a tiered system; students and other general users in the respective communities identify folksonomic terms and useful resources as "suggestions" (lo-fi), which are then winnowed by domain specialists, approved, and finally become part of the canon of knowledge (hi-fi) in the hands of more expert catalogers. Hopefully this scheme would imbue metadata and other types of literacy in general users, scholars, practitioners and library professionals, and foster the creation of metadata in regions with eGranaries that critically need it so that more information can be found. As well, it is expected to reveal culturally congruent metadata that external agents can adopt and employ. The Lo-Fi to Hi-Fi metadata pyramid can also, if successful, be globally applied to other collections and digital libraries in communities facing similar obstacles; as such obstacles are fairly universal. Before delving into the method and model, it is imperative to go over definitions of terms and provide context for the problem.

1.1. Definitions of the terms

Information literacy is defined as "the set of skills needed to find, retrieve, analyze, and use information." Those who are information literate "have learned how to learn" and find information for virtually any task ("Introduction to Information Literacy," n.d.). Computer literacy is a term being continually redefined, but Childers writes that "a person is either computer literate or not based on how proficient they are at some basic computer tasks" ("Computer Literacy," n.d.). Finally, metadata literacy is a term coined by Erik Mitchell and concerns a person's ability to cultivate adequate metadata for digital objects (Mitchell, 2010).

1.2. Computer and information literacy in areas of sub-Saharan rural Africa

For members of communities around the world, computers are critical in terms of cultivating skills necessary to be an active, participatory member of the information age. In fact, the computer as a beacon of hope and its ability to revolutionize and improve many facets of an African citizen's life was recognized in the early 1990's by Oduaran and others, but a lack of computer literacy persists even today (Oduaran, 1991; du Plessis & Webb, 2012). This paucity of computer literacy begets information illiteracy, a problem pervading not just the general, indigenous rural populace in sub-Saharan Africa but the population of teachers and information workers as well (Jager & Nassimbeni, 2007) . This problem manifests itself in not only libraries but also in the issues for which information professionals are to provide information, such as the AIDS epidemic and prevention. Compounding a lack of computer and information literacy in certain African libraries is a lack of metadata literacy, a skill not possessed even by many American library professionals (Park, Tosaka, Maszaros, & Lu, 2010).

2. WiderNet@UNC: Bringing information and literacies to the masses

2.1. Overview of WiderNet

WiderNet@UNC is a research program at the University of North Carolina, Chapel Hill that focuses on low-cost, high-impact uses of ICT and training modalities for under resourced communities worldwide. Its sister organization, the WiderNet Project, is a non-profit service program founded in 2001 that aims to bring educational digital content to places worldwide that lack adequate Internet connectivity. Using massive hard disks of material that mirror thousands of World Wide Web sites, WiderNet's eGranary Digital Libraries are currently used in over 800 locations and contain 35 million items each.

2.2. Metadata Principles and Challenges

Utilizing Dublin Core and Library of Congress (LoC) standards, WiderNet cataloguers have developed a protocol for adding metadata, highlighting resources, and creating user-centric collections, for e-Granaries. However, only a fraction of the items are findable through the catalogue because the onus falls on the small team of student and volunteer cataloguers to create records. Many more resources could be found and privileged if there was more metadata available, and if users and library workers in Africa were contributing to the process.

WiderNet has worked with partners in developing countries to create custom user-centric "portals" from catalogued records. For example, in 2008 they launched collaboration with the medical college at the University of Zambia and the School of Public Health at the University of Alabama to create a portal for teaching health sciences in Zambia. Over 1.5 million documents were garnered from the inputs of dozens of educators and practitioners around the world (lo-fi) and then WiderNet librarians cataloged over 2,000 items that had been highlighted by the expert advisors. Then, in consultation with their Zambian counterparts, they mapped 600 cataloged items to the Zambia national medical curriculum. Students and instructors were quick to adopt this curated collection and eventually insisted on it being installed in dozens of other institutions where they practiced and taught.

In another example, they worked with the United States International Council on Disabilities and over 100 advocacy groups around the world to create a dozen portals around disability rights and resources for persons with disability. Over 2.5 million new resources were added to the eGranary library and mostly librarians in the U.S. and Europe catalogued 4,000 items.

3. Hi-Fi/Lo-Fi Prototyping: Can the principle be adapted to metadata?

A prototype is defined by Merriam-Webster as an "original or first model of something from which other forms are copied or developed", or a "first or early example that is used as a model for what comes later" ("prototype", n.d.). It is proposed that methods involving high and low fidelity prototyping (hereafter called "hi" and "lo" fi) be used as a model for creating and curating metadata for resources in eGranaries. Egger describes prototypes thus:

> Low-fidelity (lo-fi) prototyping is characterized by a quick and easy translation of high-level design concepts into tangible and testable artifacts. Lo-fi is also know as low-tech, as the means required for such an implementation consist, most of the time, of a mixture of paper, cardboard, post-it notes, acetone sheets etc. A clear advantage of lo-fi prototyping is its extremely low cost and the fact that non-programmers can actively be part of the idea-crystallization process.
>
> At the other extreme, high-fidelity (hi-fi) prototypes are characterised by a high-tech representation of the design concepts, resulting in partial to complete functionality. High-tech, however, implies higher costs, both temporal and financial, and necessitates good

> programming skills to implement the prototype. The main advantage of hi-fi, high-tech prototyping is that users can truly interact with the system, as opposed to the sometimes awkward facilitator-driven simulations found in lo-fi prototyping. Obviously, there is a continuum from low to high-fidelity prototyping that usually stretches out from early to late design.("Lo-Fi vs. Hi-Fi Prototyping," n.d.)

Lo-fi prototyping, which Egger explains is "cheap, fast and accessible to non-programmers" aids participants of all levels of computer and information literacy to assist in idea and product formation, and is therefore proposed as the first step of the pyramid process, outlined in section 4.

4. The Lo-Fi to Hi-Fi Metadata Model: Crowd-Cataloguing the eGranary

This section introduces the model that is being developed, a study that is a result of meetings and the exchange of ideas at UNC Chapel Hill.

4.1. Tier 1: Lo-Fi

(lo-fi): Crowdsourcing a folksonomy

Example of Participants of Tier 1 populace: Mitchell's metadata literacy study focused on the ability of college students to create and curate metadata (Mitchell, 2010). It therefore is proposed here that college students form the majority of the lower tier of the model, the "lo-fi" stage of the process. Additionally, a machine algorithm will automatically extract metadata (indicating anything from whether something is, say, a book or web site only, to other technical details) and will feed it into this tier (or higher tiers, if a resource already contains adequate metadata to go straight to tier 2 or 3). African and international university students (graduate and undergraduate) familiar with a particular domain, e.g. hydrology, and possessing some degree of and or aptitude for metadata literacy, will create metadata. Here, terms and relationships can be drafted, thrown out and drafted again in iterative, rapid succession, in either an analog or digital environment. An eGranary resource page might have for instance a pop-up that allows one to easily tag it with descriptive terms. Alternatively, there could be paper-only environment in which students, some of whom might be more comfortable with lower-tech, are collaboratively brainstorming terms on post-its, which are later added digitally to the system. Creating terms in this manner prevents what Egger calls "tunnel vision," when people get caught up in the design of the product or resort to processes most comfortable to them, instead of focusing on what best benefits end users ("Lo-Fi vs. Hi-Fi Prototyping," n.d.). Further, people at this level are imbued with metadata, information and computer literacy through their efforts. Items in the "lo-fi" tier are not hardcoded into the canon, but Tier 1 products are passed to Tier 2 upon completion.

4.2. Tier 2: Middle-Fi

(middle-fi): Refining the terms and their relationships (synonyms, broader, narrower, if applicable). The participants are regional and international domain experts (e.g. hydrology professors/researchers, practicing hydrologists); here, the participants are fewer than in Tier 1.

4.3. Tier 3: High-Fi

(hi-fi): The smallest tier. Information specialists (African and international) approve and refine terms and relationships and add them to the canon of knowledge (hard coded as a nearly final product) in the form of a vocabulary, ontology or descriptive metadata applied to records. Domain experts and other information specialists can review this almost-final product though changes to the canon are harder to make. It is expected that at this level terms are more or less definitive and reflect what are used in a particular culture and discipline. Such high-level activities also enhance the indigenous library worker's multiple literacies, so the benefits of this process are multitudinous.

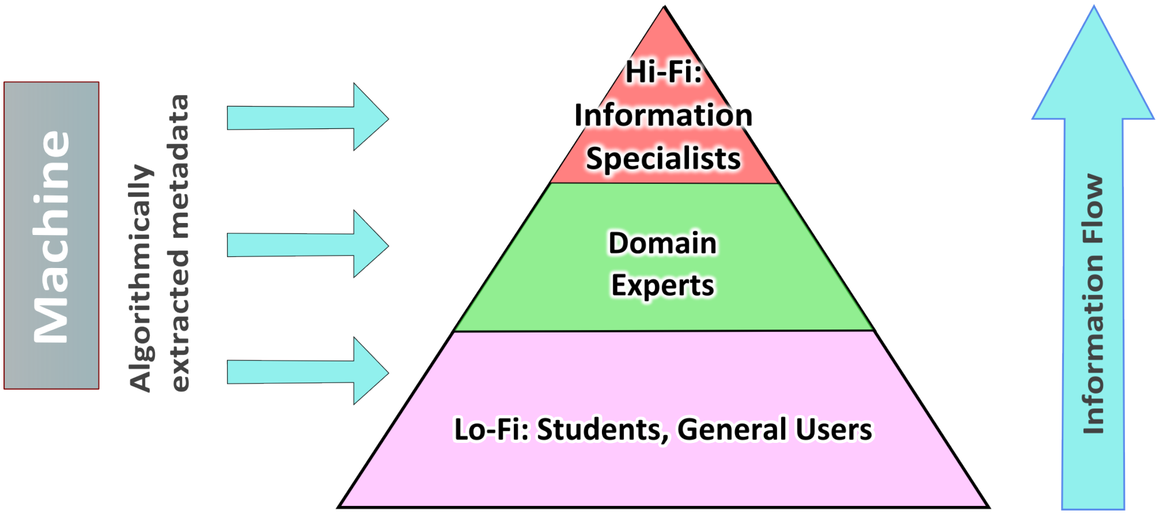

Machine
Algorithmically extracted metadata
Hi-Fi: Information Specialists
Domain Experts
Lo-Fi: Students, General Users
Information Flow

FIG. 1. Illustration of the pyramid model

TABLE 1: Summary of the pyramid model

Tier level	Type of terms	Creator population	Metadata Creator	Fidelity
1	Folksonomy, technical metadata	large	Student, machine	low
2	Folksonomy/ontology/vocabulary/technical metadata	medium	Domain expert, machine	med.
3	Ontology/vocabulary/other descriptive metadata/technical metadata	small	Information specialist, machine	high

Employing rudimentary examples of Lo-Fi/Hi-Fi metadata creation, WiderNet@UNC has demonstrated promising ideas for scaling up the creation of culturally-congruent metadata and user-centric portals through crowd-cataloging and tiered expertise. The authors will continue to explore these concepts as they expand metadata knowledge and use in target populations.

5. Conclusions

Metadata developments have progressed at a tremendous pace, particularly in technology-rich first world nations. The attention to metadata has been basic in developing countries, given more substantial priorities, such as implementing networking capacities. As technologies and opportunities such as the eGranary Digital Library are implemented, the need to address metadata issues has become increasingly apparent. This paper reported on steps taken to address metadata challenges and advance current practices. The Lo-Fi to Hi-Fi metadata pyramid model, taking its cues from other fields like design, is guiding a developing metadata initiative being pursued with the eGranary Digita Library and helping the initiative to understand how to expedite the creating of good quality metadata, making resources more findable and usable.

Next steps including testing the model in information and technology-poor areas of Africa by assessing the needs and available manpower to source the effort through a series of methods including surveys and experiments. We hope to discover through our research how to best implement the pyramid model, thereby eliminating much of the information, computer and metadata illiteracy plaguing certain areas while bolstering eGranary resource findability. If successful, the effort can be duplicated in other countries, such as Bangladesh and India, and environments such as prisons, with eGranaries.

References

Computer Literacy: Necessity or Buzzword? (n.d.). Retrieved May 16, 2014, from http://www.ala.org/lita/ital/22/3/childers

Du Plessis, A., & Webb, P. (2012). Teachers' Perceptions about their Own and their Schools' Readiness for Computer Implementation: A South African Case Study. Turkish Online Journal of Educational Technology - TOJET, 11(3), 312–325.

Global access to aging information and the gerontology healthy ageing portal. Lisa E Skemp, Ji Woon Ko, Cliff

Missen, Diane Peterson. The University of Iowa College of Nursing, Iowa City, IA, USA.

Introduction to Information Literacy. (n.d.). Retrieved May 16, 2014, from http://www.ala.org/acrl/issues/infolit/overview/intro

Jager, K. de, & Nassimbeni, M. (2007). Information Literacy in Practice: engaging public library workers in rural South Africa. IFLA Journal, 33(4), 313–322. doi:10.1177/0340035207086057

Journal of Gerontological Nursing 01/2011; 37(1):14-9.

Lo-Fi vs. Hi-Fi Prototyping: how real does the real thing have to be? (n.d.). Telono. Retrieved from http://www.telono.com/en/articles/lo-fi-vs-hi-fi-prototyping-how-real-does-the-real-thing-have-to-be/

Mitchell, E. T. (2010). Metadata literacy: An analysis of metadata awareness in college students (Ph.D.). The University of North Carolina at Chapel Hill, United States -- North Carolina. Retrieved from http://search.proquest.com.libproxy.lib.unc.edu/docview/304160575/abstract?accountid=14244

Oduaran, A. (1991). The Computer Revolution and Adult Education. Growth Prospects in Africa.

Park, J., Tosaka, Y., Maszaros, S., & Lu, C. (2010). From Metadata Creation to Metadata Quality Control: Continuing Education Needs Among Cataloging and Metadata Professionals. Journal of Education for Library & Information Science, 51(3), 158–176.

prototype. (n.d.). Merriam-Webster. Retrieved May 17, 2014, from http://www.merriam-webster.com/dictionary/prototype.

Zambia (n.d.). Sparkman Center for Global Health brings eGranaries to Universities in Zambia. Retrieved September 18, 2014, from http://www.widernet.org/node/891 and http://www.sparkmancenter.org/egranary.

How Descriptive Metadata Changes in the UNT Libraries' Collections: A Case Study

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

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

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

Daniel Alemneh
University of North Texas Libraries, USA
daniel.alemneh@unt.edu

Shadi Shakeri
University of North Texas, USA
shadishakeri@my.unt.edu

Abstract

This paper reports results of an exploratory quantitative analysis of metadata versioning in a large-scale digital library hosted by University of North Texas. The study begins to bridge the gap in the information science research literature to address metadata change over time. The authors analyzed the entire population of 691,495 unique item-level metadata records in the digital library, with metadata records supplied from multiple institutions and by a number of metadata creators with varying levels of skills. We found that a high proportion of metadata records undergo changes, and that a substantial number of these changes result in increased completeness (the degree to which metadata records include at least one instance of each element required in the Dublin Core-based UNTL metadata scheme). Another observation of this study is that the access status of a high proportion of metadata records changes from hidden to public; at the same time the reverse process also occurs, when previously visible to the public metadata records become hidden for further editing and sometimes remain hidden. This study also reveals that while most changes -- presumably made to improve the quality of metadata records -- increase the record length, surprisingly, some changes decrease record length. Further investigation is needed into reasons for unexpected findings as well as into more granular dimensions of metadata change at the level of individual records, metadata elements, and data values. This paper suggests some research questions for future studies of metadata change in digital libraries that capture metadata versioning information.

Keywords: metadata quality; distributed digital libraries; metadata change; measurement; quality assessment; best practices

1. Introduction and Background

Maintaining usable digital libraries requires high-quality metadata; one related piece involves looking at how metadata records change to determine how frequently records are edited, and, ultimately, if they have been improved. These measurements can factor into various kinds of evaluations including aspects of quality, such as "completeness," one commonly-accepted quality criterion (Moen, Stewart, & McClure, 1998; Park & Tosaka, 2010; Zavalina, 2011, etc.). Metadata completeness is evaluated as an extent to which objects are described using all applicable metadata elements to their full access capacity (Park, 2009). Of the three major metadata quality criteria (completeness, accuracy, and consistency), accuracy is the most subjective and therefore difficult to measure, while the consistency and especially completeness criteria lend themselves to variety of analyses including computational.

Stvilia and colleagues (Stvilia et al., 2004; Stvilia & Gasser, 2008) concluded that metadata changes made to improve metadata quality should be quantified and justified based on changes of value and cost of metadata to assist metadata specialists in optimizing quality assurance processes and to provide justification for spent resources. However, the analysis of literature demonstrates little research into metadata change in information science literature.

To the end of our knowledge, none of the published metadata quality studies measured metadata change. One exception is a small-scale component examining metadata change in the broader study of collection-level metadata quality in the IMLS DCC aggregation. As part of this study, researchers conducted longitudinal analysis of the modifications that had been made by digital collection developers housed at various cultural heritage institutions throughout the United States to collection-level metadata records created by hosting institutions' staff in the IMLS DCC (Zavalina, Palmer, Jackson, & Han, 2008) and found that the data values associated with the Dublin Core Collections Application Profile's Subject, Audience, Size, Spatial Coverage and Temporal Coverage metadata elements are modified the most frequently.

A number of information science studies relied on Wikipedia's so called "revision metadata" that documents who made a particular revision to the Wikipedia article and when, as well as "rollbacks" -- the process of restoring a database or program to a previously defined state -- to detect vandalism (e.g., West, Kannan, & Lee 2010; Alfonseca, Garrido, Delort, & Peñas, 2013). Similarly, Yan and McLane (2012) discussed the metadata management process for "revision metadata," including the edits, history, and tracking, made to spatial data and GIS (Geographic Information System) map figures. While using administrative metadata that documents revisions as a tool to answer other research questions, none of these studies focused on the changes made to metadata per se as opposed to information objects (e.g., Wikipedia articles) themselves.

Outside of the information science field in general and the metadata quality area in particular, one can see discussion of change in relation to texts, strings, files, etc., however, a review of the literature identified a gap in information science research in relation to analysis of metadata change. In particular, no studies to date have attempted to measure metadata change in digital libraries. The authors of this paper believe that metadata change can and should be viewed as one of the indicators of metadata quality and therefore should be examined as a step toward improving the quality of metadata in digital libraries. To begin bridging this gap, the study reported in this paper sought to answer the following research question: What is the amount of change in metadata?

The authors of this paper selected as the target for their research the centralized digital library hosted by the University of North Texas Libraries, consisting of multiple collections with varying subject scope, material types, etc. The UNT digital collections include the UNT Digital Library (containing items owned by UNT and the output of the University's research, creative, and scholarly activities), The Portal to Texas History (containing historical materials owned by more than 200 partner institutions across the state of Texas), and the Gateway to Oklahoma History (containing primarily newspapers and photographs through partnership with the Oklahoma Historical Society). The collections incorporate different types of materials including photographs, theses and dissertations, newspapers, artwork, performances, musical scores, journals, government documents, rare books and manuscripts, and posters. All items in the UNT digital collections are described using a locally-modified Dublin Core metadata schema. The digital library's infrastructure has been established according to open-source components and standards, protocols, and formats. At the time of data collection (April 18, 2014), this large-scale digital library held 691,495 unique objects, with item-level metadata records written by a number of metadata creators with varying levels of metadata creation skills.

These records reside in the digital library infrastructure operated by the UNT Libraries that is a purpose-built system for managing and providing long-term access to digital resources. Aubrey, the system used for this management, was put into production during June of 2009. The UNT Libraries placed the current metadata editing component into service in September 2009; as part of metadata management, this component versions metadata records each time they change in the system. This provides a unique collection of rich data for analysis into metadata changes.

2. Methods

According to Ochoa and Duval (2009), most of the metadata quality studies involve manual content analysis on statistically-significant samples of metadata records. Collection-level metadata records that describe entire collections of information objects as a whole, as opposed to individual objects, can still often be examined manually due to the reasonable numbers of metadata records in each sample. However, with the rapid growth of digital libraries and repositories that aggregate hundreds of thousands and often millions of items and their respective item-level metadata records, the evaluation of much more numerous item-level metadata will need to rely -- at least in part -- on computational approaches.

The study reported in this paper adopted the semi-automated quantitative research approach to analyze the entire population of metadata records in the target centralized digital library with the purpose to answer the following research question: What is the amount of change in metadata? The following broad indicators of metadata change were selected:

- frequency distribution of the number of editing events per record (i.e., how many records were edited only once and how many were edited 2, 3, or more times),
- frequency distribution of the number of editors per record,
- frequency distribution of the record length change in the process of editing,
- frequency distribution of change in record completeness (in terms of the number of metadata elements, including required elements, used), and
- frequency distribution of change in the record status (i.e., availability for the user) through the process of editing.

To measure these indicators, metadata records from the UNT Digital Library, The Portal to Texas History, and the Gateway to Oklahoma History were extracted (Phillips, 2014). The authors wrote a Python script to extract and aggregate statistics about each metadata record version into a tab-delimited format that presents a less complex view of the data (see the Appendix for the full list of data collected for each record). The dataset extraction script processed each of the 1,193,813 record instances -- including all versions of each unique record -- in the Aubrey system and calculated the number of instances (presented in the dataset as an integer) for each of the elements in the UNTL metadata scheme (UNT Libraries, 2014). Additionally, the script extracted important creation information for each metadata record including the timestamp for when it was created and last updated, the metadata creator and the last metadata modifier, whether the record is hidden to the public or unhidden, and the number of seconds that elapsed between the metadata record creation date and the metadata record edit date.

There are three fields in the dataset which may need additional description: the completeness metric, the record_length, and the record_content_length. The completeness metric calculates how "complete" a metadata record is in terms of the UNTL metadata scheme. This metric is calculated by examining the record and the existence of values for the seven fields required in our database: title, description, language, resource type, format, collection, institution, and subject. The existence or nonexistence of these values is used in a calculation that results in a number between 0 and 1, where 0 indicates a severely incomplete record with none of the required elements present, and 1 represents a complete record that has at least one instance of each of the seven elements that are required in the UNTL metadata scheme. The record_length measurement is the total number of bytes that the metadata record occupies on disk, and the record_content_length is the number of bytes of the record excluding metadata elements -- field names, qualifiers, attributes, and attribute values -- which results in the total length of data values in these metadata fields. By removing the text of metadata elements, administrative changes to the record status -- such as hiding and unhiding the record -- are not included, so a better sense of the records' full size can be seen.

3. Findings

In the dataset used for this study there are a total of 1,193,813 record instances of edited or unedited metadata record versions (see Table 1). These record instances represent 691,495 unique objects in the UNT digital collections; in the following analyses, this number is used as the "total" number of unique records in the system. The data presented in Table 1 demonstrates the steady growth in both the total number of metadata records in the system and the number of metadata records edited each year, with the highest proportion of metadata records (24.5%) added or edited in 2013.

TABLE 1: Valid edited and unedited record instances by year*.

Year	New Record Instances	Percent of Dataset
2004	928	0.1%
2005	43,425	3.6%
2006	33,899	2.8%
2007	31,053	2.6%
2008	25,138	2.1%
2009	88,580	7.4%
2010	179,498	15.0%
2011	188,810	15.8%
2012	248,439	20.8%
2013	292,342	24.5%
2014	61,695	5.2%

*Note: 6 records in the dataset are missing a metadata creation date.

As of April 2014, there were 502,675 instances of edited record versions. These versions represent 271,754 unique metadata records that have undergone changes since September 2009, when we started versioning metadata (see Table 2), or 39.3% of all metadata records in the system. Additionally, the data indicates that 9,830 records were edited one or more times before the migration to the Aubrey system and have not been edited since. That means that a total of 42.5% of all item-level metadata records in the UNT digital collections have been edited at least once. However, the records last edited before September 2009 are excluded from the edit analysis since only one -- the most current -- version of each record was retained prior to migration.

TABLE 2: Valid instances of edited records (versions) by year, September 2009-April 2014.

Year of Last Edit	Record Instances	Percentage of Edited Record Instances
2009	20,314	4.0%
2010	39,817	7.9%
2011	105,465	21.0%
2012	124,041	24.7%
2013	188,652	37.5%
2014	24,386	4.9%

The data presented in Table 2 demonstrates the steady growth in the number of metadata records edited each year, with the sharp spike (from 7.9% to 21%) in 2011 and the highest proportion of metadata records (37.5%) edited in 2013.

To get a better sense of the scope of editing frequency across the collections, we analyzed the number of edits per record and the number of editors per record. Of the edited records, nearly all (99%) have been edited five or fewer times (see Table 3), although some outlying records have been edited more than 50 times. Additionally, the majority of edited records (93.6%) have only been changed by one or two different editors (see Table 4).

For the following data analyses, edit events are compared across the entire collection of unique metadata records (n=691,495), or across the unique metadata records that have been edited at least once since September 2009 (n=271,754).

TABLE 3: Number of edits per record (n=691,495).

Number of Edits	Number of Records	Percentage of Edits	Cumulative Percentage of Edits
0	419,741	60.7%	60.7%
1	152,900	22.1%	82.8%
2	66,236	9.6%	92.4%
3	27,983	4.0%	96.4%
4	12,004	1.7%	98.1%
5	4,944	0.7%	98.8%
6	2,925	0.4%	99.2%
7	1,963	0.3%	99.5%
8	950	0.1%	99.6%
9	664	0.1%	99.7%
10	373	0.1%	99.8%
11-20	772	0.1%	99.9%
21-50	33	0.0%	100.0%
51+	7	0.0%	100.0%

TABLE 4: Number of metadata editors per record (n=271,754).

Number of Editors	Number of Records	Percentage of Records
1	197,358	72.6%
2	57,068	21.0%
3	15,397	5.7%
4	1,731	1.0%
5	180	0.1%
6	75	0.0%
7	3	0.0%
8	0	0.0%
9	0	0.0%
10	1	0.0%

In order to understand how records change over time, the authors investigated how the size of a metadata record changes during its life using the record_content_length field. The instance of this value from the first stored record (either newly created or migrated from the previous system) was compared to the most recent version in the dataset. This resulting number was categorized as an increase, a decrease, or no change in the size of the record over its life. Records that have not yet been edited have "no change." Across the entire collection, more than sixty-six percent of the records have not changed in length (see Table 5); however, among the subset of records that have been edited, more than half increased in size (see Table 6).

TABLE 5: Change in size of metadata records September 2009-April 2014 (n=691,495).

Change Category	Number of Records	Percentage of All Records
No Size Change (0)	459,350	66.4%
Size Increase (+)	146,046	21.1%
Size Decrease (-)	86,099	12.5%

TABLE 6: Change in size of edited metadata records September 2009-April 2014 (n=271,754).

Change Category	Number of Records	Percentage of Edited Records
No Size Change (0)	39,610	14.6%
Size Increase (+)	146,046	53.7%
Size Decrease (-)	86,099	31.7%

The authors took a similar approach to determine the change in completeness among records across time as they did for calculating the record content length over time (using the automatically-calculated metric that measures the presence of all required fields in a metadata record). The earliest value of completeness from the record samples was compared with the most recently edited values to determine whether the completeness increased, decreased, or stayed the same. A large majority of the whole collection -- nearly 96% -- had no change in completeness (see Table 7); and, even among the subset of edited records, roughly 90% had no change in completeness (see Table 8). Overall, completeness generally stayed the same or increased, although thirteen records decreased in completeness, likely due to a mistake or misunderstanding when editing.

TABLE 7: Change in completeness of metadata records September 2009-April 2014 (n=691,495).

Change Category	Number of Records	Percentage of All Records
No Completeness Change (0)	662,508	95.8%
Completeness Increase (+)	28,974	4.2%
Completeness Decrease (-)	13	0.0%

TABLE 8: Change in completeness of edited metadata records September 2009-April 2014 (n=271,754).

Change Category	Number of Records	Percentage of Edited Records
No Completeness Change (0)	242,767	89.3%
Completeness Increase (+)	28,974	10.7%
Completeness Decrease (-)	13	0.0%

Aside from general size and completeness of records, the final research indicator involves an aspect of particular interest in this analysis, which relates to the accessibility of records to the public. In UNTL metadata, records contain a field that controls whether or not a record is hidden; if the value is "true," the record cannot be viewed in any way without administrative access to the item. For items that have a hidden value of "false," the metadata record is visible to the public and searchable. This value only governs the metadata record and does not affect the accessibility of the item (i.e., items that have restricted usage or embargoes can still have a hidden value of "false").

First, to see how this value changes over time, the authors compiled statistics for the number of records for which the record access status value has changed -- either hidden to visible, or visible to hidden. More than eighty percent of unique metadata records in the system have not changed in access status (see Table 9), while a lesser majority (65%) of the edited records remained unchanged (see Table 10).

TABLE 9: Change in access status of metadata records September 2009-April 2014 (n=691,495).

Change Category	Number of Records	Percentage of All Records
Access Status Changed	94,516	13.7%
Access Status Unchanged	596,979	86.3%

TABLE 10: Change in access status of edited metadata records September 2009-April 2014 (n=271,754).

Change Category	Number of Records	Percentage of Edited Records
Access Status Changed	94,516	34.8%
Access Status Unchanged	177,238	65.2%

In general, looking at how record access status has changed is important since it affects accessibility and usage, however, we particularly want to highlight records that have moved from a visible status to a hidden status. This event represents a situation in which a digital object that was available to the public -- and may have been viewed, cited, or linked -- is no longer available. Tables 11 and 12 present a more detailed analysis of this kind of metadata change, breaking down the number of records that had a value of "false" (visible) that changed to "true" (hidden) at any point in their edit history. For comparison, Tables 11 and 12 also contain statistics for records that did not change access status, but an additional column gives the current status of each set of records, providing detail as to how many records are unchanged but visible, versus unchanged but hidden.

Overall, more than ninety percent of the all metadata records currently have a hidden value of "false," making them publicly accessible (see Table 11). More than 60% of the records that have been edited have started as visible and not changed, while another 33% have been changed in access status from hidden to visible during the course of editing (see Table 12).

TABLE 11: Current (April 2014) access status and status changes across all records (n=691,495).

Change Category	Changed from Visible to Hidden	Final Hidden Value	Number of Records	Percentage of All Records
Access Status Changed	No	False (Visible)	90,295	13.1%
Access Status Changed	Yes	False (Visible)	1,899	0.3%
Access Status Changed	Yes	True (Hidden)	2,322	0.3%
Access Status Unchanged	No	False (Visible)	553,262	80.0%
Access Status Unchanged	No	True (Hidden)	43,717	6.3%

TABLE 12: Current (April 2014) access status and status changes across edited records (n=271,754).

Change Category	Changed from Visible to Hidden	Final Hidden Value	Number of Records	Percentage of Edited Records
Access Status Changed	No	False (Visible)	90,295	33.2%
Access Status Changed	Yes	False (Visible)	1,899	0.7%
Access Status Changed	Yes	True (Hidden)	2,322	0.9%
Access Status Unchanged	No	False (Visible)	167,478	61.6%
Access Status Unchanged	No	True (Hidden)	9,760	3.6%

The rows that have particular significance in Tables 11 and 12 show statistics for the records that have changed in status from visible to hidden at some point in their history. Forty-five percent of those 4,221 records have ultimately been edited in some way and then made visible again. However, the other fifty-five percent (2,322 records) have remained hidden and may need additional review.

4. Discussion and Conclusions

In summary, the data in this paper outlines information to answer some general questions about change in metadata records across a body of digital items, as a preliminary step toward further research. This study revealed that a high proportion of metadata records in the UNT digital collections (almost 40%) have been edited at least once in the period between September 2009 and April 2014 to change record content and/or access status. In addition, our data provides evidence that the purposive metadata change activity -- expressed in the sheer number and proportion of edited records -- has steadily and substantially grown over time. These findings support the assumption that metadata is a constantly-evolving resource.

Several other points particularly stood out as part of this analysis. First, a considerable number (nearly 11%) of edited records improved in quality based solely on the "completeness" metric. Although this does not give a holistic view of the final metadata quality of those records (in particular, with regards to accuracy, consistency or record completeness beyond the mere presence of at least one instance of each required metadata element), in general, metadata editors are adding required information when it is missing, improving the overall value of the metadata.

Next, regarding change in length, a larger than expected number of edited records (31.7%) decreased in size as a result of changes, suggesting the removal of information. However, since the record_content_length indicator represents the total number of characters in the record, even minor changes could have accounted for a net decrease in record length, such as the removal of an extra space, the correction of typographical errors with extra letters/characters, or the replacement of longer placeholder values with shorter actual values as editors completed partial records. Additionally, qualifiers and terms from controlled vocabularies contribute to the length, so changing those values could decrease the number of characters. Based on this understanding, a decrease in record length does not necessarily equate to a loss of information, or a decrease in the quality or accuracy of a particular record.

Finally, as noted in the previous section, a number of metadata records (2,322) were hidden at the time of data collection, even though they had been visible at some point in their edit history. Although it is a small subset within the whole system -- only .3% of the total records -- any links to those records have been broken. Since the general goal is to provide as much access as possible and maintain permanent links to items and their respective metadata records in the UNT digital collections, those records should be reviewed to see if changes would allow them to become accessible once again, and to gain details about the circumstances in order to limit or avoid similar situations in the future.

4.1. Further Study

The research reported in this paper is a case study that sought to explore quantitative dimensions of metadata change and its general effects within a large digital collection. It helps identify some areas for future exploration that will be addressed by further, more in-depth, mixed-methods studies. These future studies will need to examine both quantitative and qualitative characteristics of metadata change in various digital repositories to answer these and other research questions:

- What is the frequency of change? What is the distribution of the lengths of time between initial record creation and its first modification; between the first and subsequent modifications?
- How does the number of instances of key metadata elements (such as title, creator, description, subject, etc.) change in the process of editing?

- Which common metadata change categories can be identified? What is the relative frequency of occurrence of these metadata change categories?
- Which elements in metadata records are changed the most often?
    - How do they change?
    - How do these changes affect the overall quality – completeness, consistency, and accuracy – of metadata records?

To answer these and other more specific research questions, future studies will need to involve in-depth manual comparative analysis of versions for a manageable sample of metadata records. The role of the current exploratory study is to serve as the first stepping stone and to spur interest among metadata practitioners in conducting research into metadata change.

With major digital content management tools (e.g., Fedora, Islandora, and Hydra) now incorporating metadata versioning, more and more digital repositories will be able to capture versions of their metadata records and explore the change in their metadata over time. Further work by other institutions in this same area could allow for important comparative research. Without similar data, there is no way to evaluate whether the findings in this study are consistent across most digital libraries, or to determine the significance of any situations in which the experience at UNT differs from other digital libraries. Results of measuring metadata change will also help to determine the overall metadata quality, compare metadata quality across different collections of items, and will inform metadata management decisions such as setting priorities in metadata quality.

References

Alfonseca, E., Garrido, G., Delort, J., & Peñas, A. (2013). WHAD: Wikipedia historical attributes data. Language Resources and Evaluation, 47(4), 1163-1190. DOI: http://dx.doi.org/10.1007/s10579-013-9232-5

Moen, W.E., Stewart, E.L, & McClure, C.R. (1998). The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study. Retrieved from: http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. International Journal of Digital Libraries, 10, 67-91.

Park, J. (2009). Metadata quality in digital repositories: a survey of the current state of the art. Cataloging & Classification Quarterly, 47 (3), 213-228.

Park, J. & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. Cataloging & Classification Quarterly, 48 (8), 96-715.

Phillips, M. (April 2014). UNT Libraries metadata edit dataset. Retrieved from: http://digital.library.unt.edu/ark%3A/67531/metadc304852/

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., & Cole, T. (2004). Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. In Thompson, H.A. (Ed.). Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, pp. 223-237.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., & Cole, T. (2004). Metadata quality for federated collections. Proceedings of ICIQ04, 111-125.

Stvilia, B. & Gasser, L. (2008). Value-based metadata quality assessment. Library & Information Science Research, 30(1), 67-74.

University of North Texas Libraries (2014). Input Guidelines for Descriptive Metadata (Revised version). Retrieved from: http://www.library.unt.edu/digital-projects-unit/input-guidelines-descriptive-metadata.

West, A.G., Kannan, S., & Lee I. (2010). STiki: An anti-vandalism tool for Wikipedia using spatio-temporal analysis of revision metadata. Proceedings of the 6th International Symposium on Wikis and Open Collaboration (WikiSym '10). DOI=10.1145/1832772.1832814

Yan, Y., & McLane, T. (2012). Metadata management and revision history tracking for spatial data and GIS map figures. Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. DOI=10.1145/2345316.2345357

Zavalina, O.L., Palmer, C.L., Jackson, A.S., & Han, M.-J. (2008). Evaluating descriptive richness in collection-level metadata. Journal of Library Metadata, 8 (4), 263-292.

Zavalina, O.L. (2011). Contextual metadata in digital aggregations: Application of collection-level subject metadata and its role in user interactions and information retrieval. Journal of Library Metadata, 11(3/4), 104-128.

Appendix

Alphabetical list of information captured in the dataset for all versions of metadata records in the UNT.

Field	Example	Description
sample_id	ark:/67531/metacrs10000_2009-12-20T02:07:08	Unique identifier for a sample record version.
ark	ark:/67531/metacrs10000	Unique record identifier.
citation	0	Number of citation element entries.
collection	1	Number of collection element entries.
completeness	0.983050847458	Completeness metric.
contributor	0	Number of contributor element entries.
coverage	1	Number of coverage element entries.
creator	4	Number of creator element entries.
date	0	Number of date element entries.
degree	0	Number of degree element entries.
description	2	Number of description element entries.
format	1	Number of format element entries.
hidden	False	Record hidden status (true/false).
identifier	2	Number of identifier element entries.
institution	1	Number of institution element entries.
language	1	Number of language element entries.
meta	11	Number of meta element entries.
metadata_creation_date	2007-06-12, 16:50:25	Date and time record was created.
metadata_creator	mphillips	Username for the record creator.
metadata_edit_date	2008-02-18, 15:22:21	Date and time record was last edited.
metadata_editor	govdocs	Username of the last metadata editor.
note	0	Number of note element entries.
primarySource	0	Number of primary source element entries.
publisher	1	Number of publisher element entries.
record_content_length	1775	Record length in bytes, excluding "meta" fields.
record_length	2445	Size of the metadata record in bytes.
relation	0	Number of relation element entries.
resourceType	1	Number of resource type element entries.
rights	1	Number of rights element entries.
source	0	Number of source element entries.
subject	12	Number of subject element entries.
time_since_creation	2168116	Time in seconds from record creation to last edit.
title	1	Number of title element entries.

Metadata in Support of Research

Metadata Integration for an Archaeology Collection Architecture

Sivakumar Kulasekaran
Texas Advanced Computing Center
The University of Texas at Austin
siva@tacc.utexas.edu

Jessica Trelogan
Institute of Classical Archaeology
The University of Texas at Austin
j.trelogan@austin.utexas.edu

Maria Esteva
Texas Advanced Computing Center
The University of Texas at Austin
maria@tacc.utexas.edu

Michael Johnson
L - P : Archaeology, United Kingdom
m.johnson@lparchaeology.com

Abstract

During the lifecycle of a research project, from the collection of raw data through study to publication, researchers remain active curators and decide how to present their data for future access and reuse. Thus, current trends in data collections are moving toward infrastructure services that are centralized, flexible, and involve diverse technologies across which multiple researchers work simultaneously and in parallel. In this context, metadata is key to ensuring that data and results remain organized and that their authenticity and integrity are preserved. Building and maintaining it can be cumbersome, however, especially in the case of large and complex datasets. This paper presents our work to develop a collection architecture, with metadata at its core, for a large and varied archaeological collection. We use metadata, mapped to Dublin Core, to tie the pieces of this architecture together and to manage data objects as they move through the research lifecycle over time and across technologies and changing methods. This metadata, extracted automatically where possible, also fulfills a fundamental preservation role in case any part of the architecture should fail.

Keywords: archeology; collection architecture; metadata integration; automated metadata extraction; ARK; iRODS rules; Corral; Rodeo; Ranch

1. Introduction

Data collections are the focal point through which study and publishing are currently accomplished by large research projects. Increasingly they are developed across what we refer to as *collection architectures*, in which data and metadata are curated across multi-component infrastructures and in which tasks such as data analysis and publication can be accomplished by multiple users seamlessly and simultaneously across a collection's lifecycle. It is well known that metadata is indispensable in furthering a collection's preservation, interpretation, and potential for reuse, and that the process of documenting data in transition to an archival collection is essential to those goals. In the collection architecture we present here, we use metadata in a novel way: to integrate data across recordkeeping and archival lifecycle phases as well as to manage relationships between data objects, research stages, and technologies. In this paper, we introduce and illustrate these concepts through the formation of an archaeological collection spanning many years. We show how metadata, formatted in Dublin Core (DC), is used to bridge data and semantics developed as teams and research methods have changed over the decades.

The model we propose differs from traditional data management practices that have been described as the "long tail of research" (Wallis et al., 2014), in which researchers may store data in scattered places like home computers, hard-drives and institutional servers, with data integrity potentially compromised. Without a clear metadata strategy, data provenance becomes blurry and integration impossible. In the traditional model, archiving in an institutional repository or in a data publication platform comes at the end of the research lifecycle, when projects are finalized, often decades after they started, and sometimes too late to retain their original intended meaning (Eiteljorg, 2011). At that final stage, reassembling datasets into collections that can be archived and shared becomes arduous and daunting, preventing many from depositing data at all. Instead, a collection architecture such as the one presented here, which is actively curated by the research team throughout a project, helps to keep ongoing research organized, aggregates metadata on the go, facilitates data sharing as research progresses, and enables the curator-researcher to control how the public interacts with the data. Moreover, data that are already organized and described can be promptly transferred to a canonical repository.

For research projects midway between the "long tail" and the new data model, the challenge is to merge old and new practices, to shape legacy data into new systems without losing meaning and without overwriting the processes through which data were conceived. We present one such case: a collection created by the Institute of Classical Archaeology (ICA, 2014) representing several archaeological investigations (excavations, field surveys, conservation, and study projects) in Italy and Ukraine going back as far as the mid-1970s. As such, it includes data produced by many generations of research teams, each with their own idiosyncratic recording methods, research aims, and documentation standards. Integrating it into a collection architecture that is accessible for ongoing study while thinking ahead about data publishing and long-term archiving has been the subject of ongoing collaboration between ICA and the Texas Advanced Computing Center (TACC, 2014) for the last five years (Trelogan et al., 2010; Walling et al., 2011; Rabinowitz et al., 2013).

In this project metadata is at the center of a transition from a disorganized aggregation of data—belonging to both the long tail of research, and new data that is being actively created during study and publication—into a collection architecture. The work has involved re-engineering research workflows and the definition of two instances of the collection with different functions and structures: one is a stable collection which we call the *archival instance* and the other, *a study and presentation instance*. Both are actively evolving as research continues, but the methods we have developed allow researchers to archive data on the fly, enter metadata only once, and to move documented data from the archive into the presentation instance and vice versa, ensuring data integrity and avoiding the duplication of effort. The DC standard integrates the data objects within the collection and binds the collection instances together.

2. Archaeology as the Conceptual Framework for a Collection Architecture

Archaeology is an especially relevant domain for exploring issues of data curation and management because of the sheer volume and complexity of documentation produced during the course of fieldwork and study (Kansa et al., 2011). Likewise, because a typical archaeological investigation requires teams of specialists from a large number of disciplines (such as physical anthropology, paleobotany, geophysics, and archaeozoology) a great deal of work is involved in coordinating the datasets produced (Faniel et al., 2013). Making such coordination even more challenging is the tendency for large archaeological research projects, like those in the ICA collection, to carry on for multiple seasons, sometimes lasting for decades. Projects with such long histories and large teams can contain layer upon layer of documentation that reflect changes in technologies, standard practices, methodologies, teams, and the varied ways in which they record the objects of their particular study.

As in an archaeological excavation, understanding these sediments is key to unlocking the collection's meaning and to developing strategies for its preservation. Due to the inevitable lack of consistency in records that span years and specialties, these layers can easily become closed-off information silos that make it impossible to understand their purpose or usefulness. The work we are doing focuses on revealing and documenting those layers through metadata, without erasing the semantics of past documentation, and without a huge investment of labor at the end. To address these challenges within the ICA collection, we needed a highly flexible, lightweight solution (in terms of cost, time, and skills required to maintain) for file management, ongoing curation, publishing, and archiving.

3. Functional and Resource Components of the Collection Architecture

Currently the ICA collection is in transition from disorganized data silos to an organized collection architecture, illustrated in Figure 2. The disorganized data, recently centralized in a networked server managed by the College of Liberal Arts Instructional Technology Service (LAITS, 2014), represents an aggregation of legacy data that had been previously dispersed across servers, hard-drives and personal computers. The data were centralized there to round up and preserve disconnected portions of the collection so that active users could work collaboratively within a single, shared collection. Meanwhile, new data are continuously produced as paper records are digitized and as born-digital data are sent in from specialists studying abroad. To manage new data and consolidate the legacy collection, we created a recordkeeping system consisting of a hierarchical file structure implemented within the file share, with descriptive labels and a set of naming conventions for key data types, allowing users to promptly classify the general contents and relationships between data objects while performing routine data management tasks (see Figs. 1 and 5). The recordkeeping system is used as a staging area where researchers simultaneously quality check files, describe and organize them (by naming and classifying into labeled directories) and purge redundant copies, all without resorting to time-consuming data entry. Once organized, data are ingested into the collection's archival instance (See Fig. 2) where they are preserved for the long term and can be further studied, described, and exposed for data sharing.

3.1. Staging and recordkeeping system: gathering basic collection metadata

Basic metadata for the collection is generated from the recordkeeping system mentioned above. Using the records management big bucket theory (Cisco, 2008) as a framework, we developed a file structure that would be useful and intuitive for active and future research and extensible to all of the past, present, and future data that will be part of the ICA collection (Fig. 1). This file structure was implemented within the fileshare and is mirrored in the archival instance of the collection for a seamless transition to the stable archive. The core organizing principle for the data is its provenance as the archaeological "site" or "project" for which it was generated. Within each of these larger "buckets", we group data according to three basic research phases appropriate to any investigation encountered in the collection, be it surface survey, geophysical prospection, or excavation[1]: 1) field, 2) study, 3) publication. These top two tiers of the hierarchy allow us to semantically represent, per project, what we consider primary or raw versus processed, interpreted data, and the final polished data that are tied to specific print or online publications. The third tier includes classes of data recorded during fieldwork and study (e.g. field notes, site photos, object drawings) and the subjects of special investigations (e.g. black-gloss pottery, physical anthropology, or paleobotany). The list was generated inductively from the materials produced during specific investigations and is applicable to most ICA projects. As projects continue through the research lifecycle this list may expand to add other materials that were not initially accounted for. Curators can pick the appropriate classes and file data accordingly. Files are named according to a convention (Fig. 5), which encodes provenance, relationships between objects found together, the subject represented (e.g. a bone from a specific context), as well as the process history of the data object (e.g. a scanned photograph).

This recordkeeping system is invaluable for the small team at ICA managing large numbers of documentation objects (>50,000 per each of over two dozen field projects). Because many

[1] This is, in fact, an appropriate way to describe the lifecycle of any kind of investigation – archaeological or otherwise – that involves a fieldwork or data-collection stage.

projects in ICA's collection are still in the study phase and do not yet have a fully developed documentation system, the filenames and directories are often the sole place to record metadata. As the data are moved to the new collection architecture, the metadata is automatically mapped as a DC document with specific qualifiers that preserve provenance and contextual relationships between objects. Metadata is thus entered only once, and is carried along through the archival to the study and presentation instances where specialists may expand and further describe them as they study and prepare their publications.

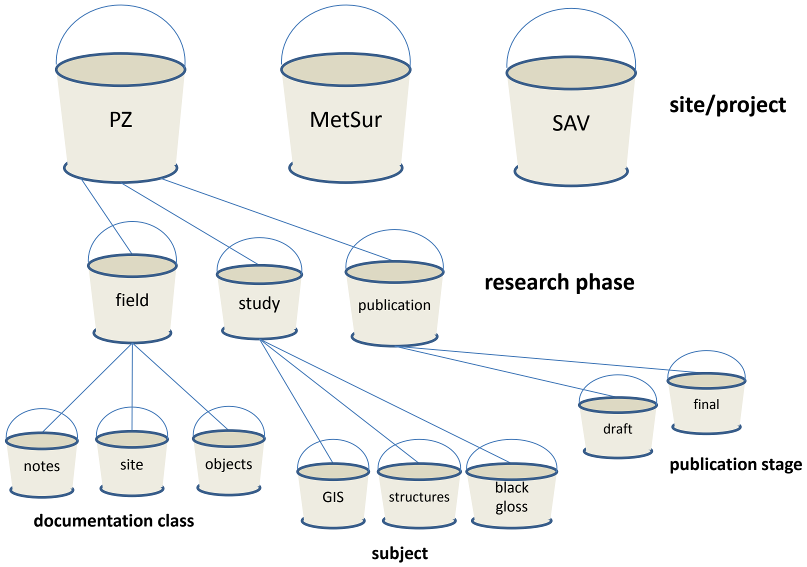


FIG. 1. The highest levels of the file structure, represented here as "big buckets" whose labels embed metadata about the project, stages of research, classes of documentation, and subjects of specialist study.

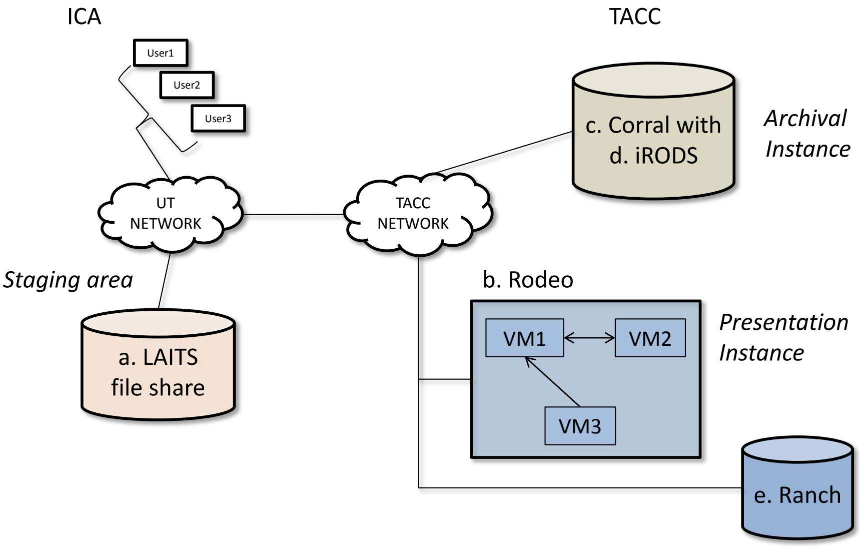


FIG. 2. Resource components of ICA's collection architecture: a. LAITS file share (staging area); b. Rodeo, cloud computing resource that hosts Virtual Machines (VMs); c. Corral, storage resource that contains active collections; d. iRODS, data management system; e. Ranch, tape archive for backups and long-term storage.

3.2. Archival instance: Corral/iRODS

Corral is a high performance resource maintained by TACC to service UT System researchers (TACC, 2014; Corral, 2014). This system includes 6 petabytes of on- and off-site storage for data replication, as well as data management services through iRODS (integrated Rule-Oriented Data System) (iRODS, 2014). iRODS is an open-source software system that abstracts data from storage in order to present a uniform view of data within a distributed storage system. In iRODS a central metadata database called iCAT holds both user defined and system metadata, and a rule engine is available to create and enforce data policies. We implemented custom iRODS rules to automate the metadata extraction process. To access the data on Corral/iRODS, users can use GUI-based interfaces like iDROP and WebDAV or a command-line utility. Data on Corral/iRODS are secured through geographical replication to another site at UT Arlington.

3.3. Presentation instance

3.3.1. ARK

To provide a central platform for collaborative study of all material from each project, to record richer descriptions and interpretations, and to define complex contextual relationships, we adopted ARK, the Archaeological Recording Kit (ARK, 2014). ARK is a web-based, modular "toolkit" with GIS support, a highly flexible and customizable database and user interface, and a prefabricated data schema to which any kind of data structure can be mapped (Eve et al., 2008). This has allowed ICA staff to create—relatively quickly and easily—a separate ARK for each site or project, and to pick and choose the main units of observation within that (e.g. the "site" in the case of a survey project, or the "context" and "finds" for an excavation project). At ARK's core are user-configured "modules", in which the data structure is defined for each project. In terms of the "big buckets" shown in Fig. 1, each of the top tier (site/project) buckets can have an implementation of ARK, with custom modules that may correspond to the documentation classes and/or study subjects represented in the third tier of buckets, depending on the methodological approach.[2] Metadata mappings are defined within the modules in each ARK (e.g., Fig. 6). This presentation instance allows the user to interact with data objects that reside in the archival instance on Corral/iRODS, describe them more fully in context of the whole collection (creating more metadata), and then push that metadata back to the archival instance.

3.3.2. Rodeo

Rodeo is TACC's cloud and storage platform for open science research (RODEO, 2014). It provides web services, virtual machine (VM) hosting, science gateways, and storage facilities. Virtual machines can be defined as a "software based emulation of a computer" (VM, 2014). Rodeo allows users to create their own VM instance and customize it to perform scientific activities for their research needs. All of the ARK services, including the front-end web services, databases, and GIS, are hosted in Rodeo's cloud environment. We use three VM instances to host each of these services. To comply with best security practices we separate out the web services from the GIS and the databases. If the web service is compromised or any security issues arise, none of the other services are affected and only the VM that hosts the affected web service needs to be recreated. During the study and publication stages, data on iRODS are called from ARK, and metadata from ARK is integrated into the iCAT database.

[2] We currently have three live implementations of ARK hosted at TACC, one housing legacy data from excavations carried out from the 1970s to the 1990s, recorded with pen and paper and film photography with finds as the main unit of observation; a contemporary excavation, from 2001 to 2007, which was mostly born digital (digital photos, total station, in-the-field GIS, etc.) and focused on the stratigraphic context; and one survey project, from the 1980s to 2007, consisting of a combination of born digital and digitized data and centered on the "site" and surface scatters of finds.

3.3.3. Ranch

Ranch is TACC's long-term mass storage solution with a high-performance tape-based system. We are using it here as a high-reliability backup system for the publication instance of the collection and its metadata hosted in Rodeo on the VMs. We also routinely back up the ARK code base and custom configurations. Across Corral and Ranch, the entire collection architecture is replicated for high data availability and fault tolerance.

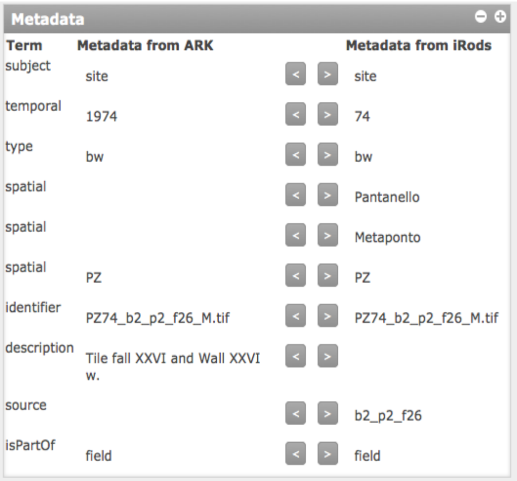
4. Workflow and DC Metadata

4.1. Automated metadata extraction from the recordkeeping system

To keep manual data entry to a minimum, we developed a method for automatically extracting metadata embedded in filenames and folders of our recordkeeping system. We used a modularized approach using Python (Python, 2014) and customized iRODS rules so that individual modules can be easily plugged in or reused for other collections. One module extracts technical metadata using FITS (FITS, 2014) and maps the extracted information to DC and to PREMIS (PREMIS, 2014) using an XSLT stylesheet. Another module creates a METS document (METS, 2014) also using a XSLT stylesheet transformation from the FITS document. The module focusing on descriptive metadata extracts information from the recordkeeping system and maps it to DC following the instructions from the data dictionary. Metadata is integrated into a METS/DC document. Finally, metadata from the METS document is parsed and registered in the iCAT database (Walling et al., 2011). Some files do not conform to the recordkeeping system because they could not be properly identified and thus named and classified. For those, the descriptive metadata will be missing and only a METS document with technical metadata is created, with the technical information added into iCAT. This metadata extraction happens on ingest to iRODS, so it occurs only as frequently as the users upload data that are understood and organized by the researchers. The accuracy of the extracted metadata depends upon the accuracy of the filenames (e.g., adherence to naming convention or correctness of object identification). These are then further quality checked within the ARK interface during detailed collaborative study, and corrections are pushed back to the iRODS database as needed by the user.

4.2. Syncing data between ARK and iRODS

The next phase was to sync metadata between the two databases: ARK and iCAT/iRODS. A new function was created within ARK to pull in metadata from iRODS and display it alongside the metadata from ARK for each item in a module (e.g. object photographs).



Term	Metadata from ARK	Metadata from iRods
subject	site	site
temporal	1974	74
type	bw	bw
spatial		Pantanello
spatial		Metaponto
spatial	PZ	PZ
identifier	PZ74_b2_p2_f26_M.tif	PZ74_b2_p2_f26_M.tif
description	Tile fall XXVI and Wall XXVI w.	
source		b2_p2_f26
isPartOf	field	field

FIG. 3 Metadata subform from ARK, allowing user to compare the information from the two collection instances.

Fields in ARK are used to define what data are stored where in the back-end ARK database, the way that they should be displayed on the front-end website, and the way that they should be added or edited by a researcher. The data classes used in ARK are specific to that environment and have been customized and defined according to user needs within each implementation. The mapping between the DC term and the corresponding field within ARK is defined in the module configuration files.

While research progresses, data and metadata are added and edited via the ARK interface. The user can update the metadata in iRODS from ARK or vice versa, using arrow buttons showing the direction that the data will move. The system automatically recognizes if the user is performing an add or edit operation. PHP is used to read and edit the information from ARK and iRODS, and Javascript is used to give the user feedback and confirm the modifications (Fig. 3). The metadata linked to either the DC term or the ARK field are then presented and updated through the ARK web interface.

The workflow represented in Fig. 4 allows us to transition data into the collection architecture and to perform ongoing data curation tasks throughout the research lifecycle. Note that in this workflow, data are ingested first to the archival instance of the collection. This allows archiving as soon as data are generated, assuring integrity at the beginning of the research lifecycle.

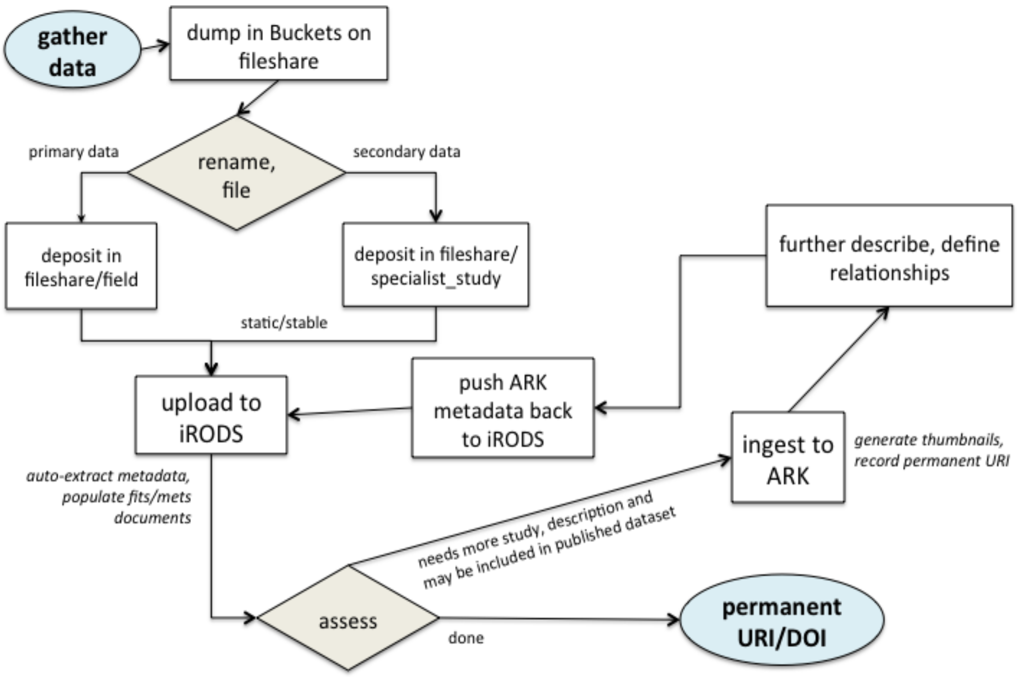

gather data
dump in Buckets on fileshare
primary data
rename, file
secondary data
deposit in fileshare/field
deposit in fileshare/ specialist_study
static/stable
further describe, define relationships
upload to iRODS
push ARK metadata back to iRODS
ingest to ARK
generate thumbnails, record permanent URI
auto-extract metadata, populate fits/mets documents
needs more study, description and may be included in published dataset
assess
done
permanent URI/DOI

FIG. 4. Curation workflow.

4.3. Dublin Core metadata: the glue that binds it all together

Metadata schemas are typically used to describe data for ease of access, to provide normalization, and to establish relationships between objects. They can be highly specialized to include elements that embed domain-specific constructs. A general schema like DC, on the other hand, can be used in most disciplines, if fine-grained description is not a priority. In choosing a schema for this project we considered its ability to relate objects to one another, its generalizability in representing the wide range of recording systems represented in the collection, and its ease of use. With this in mind, we chose to use DC, which is widely used for archaeological applications, including major data repositories like the UK-based Archaeology Data Service (ADS, 2014) and, in the US, the Digital Archaeological Record (tDAR, 2014).

In this project the DC standard is a bridge over which data are exchanged between collection instances and across active research workflows, turning non-curated into curated data, while providing a general, widely understood method for describing the collection and the relationships between the objects. Given the need for automated metadata extraction and organization processes, we required higher levels of abstraction to map between the different organizational and recording systems, data structures, and concepts used over time. Furthermore, DC is the building block for future mapping to a semantically rich ontology like CIDOC-CRM (CRM, 2014), a growing standard that is used for describing cultural heritage objects that is particularly relevant for representing archaeology data in online publishing platforms (OpenContext, 2014). CIDOC-CRM provides the scope to fully expose the richness of exhaustive analysis, and allows the precise expression of contextual relationships between objects of study, as well as the research process and events (historical and within an excavation or study), provenance (of cultural artifacts as well as of data objects), and people. Such semantic richness, however, only fully emerges at the final stages of a project, and we are here concerned with ongoing work resulting in a collection that is still in formation and evolving rapidly.

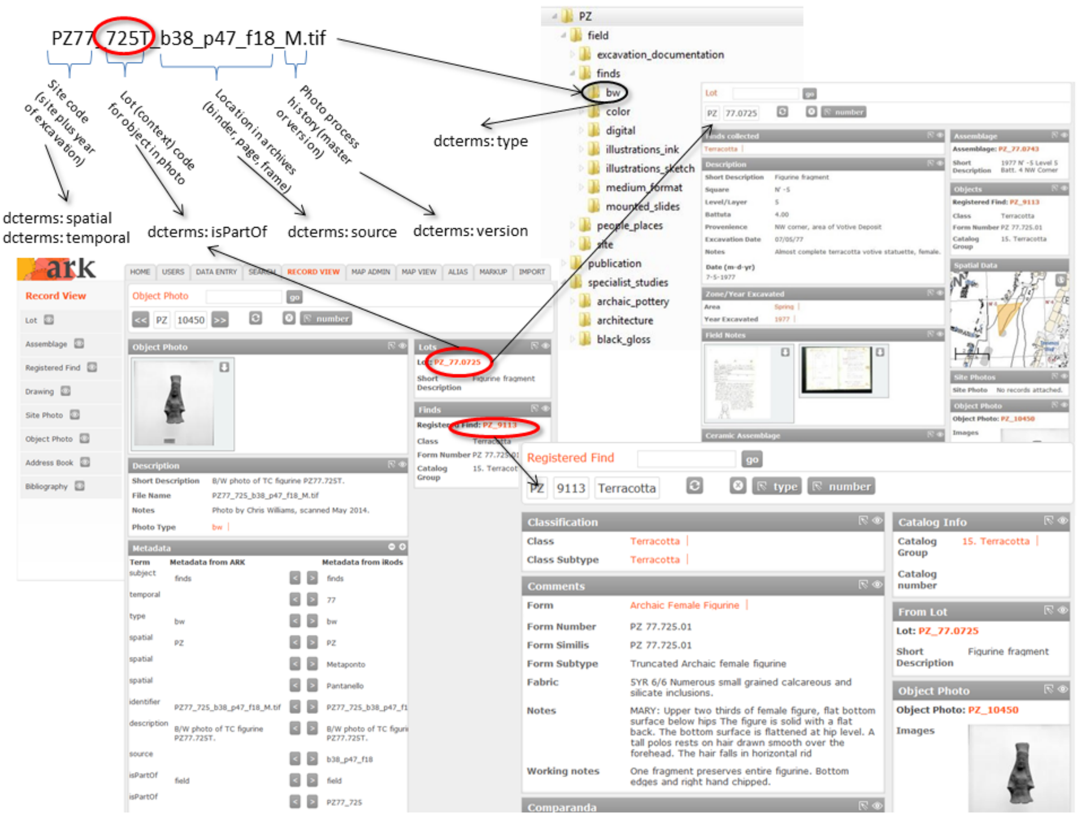


FIG. 5. Metadata extracted from filename and folder labels are mapped to DC terms. Once in ARK further descriptive metadata can be added and pushed back to iRODS.

4.4. Metadata mapping and its semantics

The mapping to DC for this project was considered in two stages. For the archival instance of the collection, we focused on expressing relationships between individual data objects (represented by unique identifiers) through the DC elements "spatial," "temporal," and "isPartof." This allows grouping, for example, of all the documentation from a given excavation, or all artifacts found within the same context. We also categorized documentation types and versions to help us relate data objects to the physical objects they represent (e.g., a drawing or photo of an artifact). For the publication instance presented in ARK, mapping focused on verbal descriptions, interpretations, and the definition of relationships produced during study. These then populate the "description" and "isPartOf" elements in the DC document. As a data object enters the collection to be further analyzed and documented in ARK, all the key documentation related to that object is exchanged over time throughout all pieces of the collection architecture and remain in the archival instance once complete. For example, when a photo is scanned, named, and stored in the appropriate folder, this embeds provenance information for the object in the photo (e.g., context code, site and year of excavation), the provenance of the photo itself (e.g., location of negative in physical archive), the process history of the data object (e.g., raw scan or an edited version), its relations to other objects in the collection, and the description created by specialists in ARK (see Fig. 5). For the data curator, the effort is minimal, and information is extracted automatically and mapped to terms that are clearly understood. The information is carried along as the data object moves from the primary data archive to the interpretation platform, and is enhanced through study and further description every time the metadata is updated. By mapping key metadata elements to DC (Table 1) we reduce data entry and provide a base for future users of the collection.

TABLE 1. Extract of a data dictionary that maps the fields in an ARK object photo module to the recordkeeping system and DC elements.

ARK term	ARK field	Record Keeping Example	DC Term
Short Description	$conf_field_short_desc	Terracotta Figurine	description
File Name	$conf_field_filename	PZ77_725T_b38_p47_f18_M.tif	identifier
Photo Type	$conf_field_phototype	PZ/field/finds/bw	format
Date Excavated	$conf_field_excavyear	1977	date
Date Photographed	$conf_field_takenon	1978	created
Photographed by	$conf_field_takenby	Chris Williams	creator
Area	$conf_field_area	Pantanello	spatial
Zone	$conf_field_zone	Sanctuary	spatial

4.5. Metadata for integrity

In addition to the technical metadata extraction, descriptive metadata added throughout the research lifecycle assures the collection's integrity in an archaeological sense by reflecting relationships between data objects. Moreover, because we have the same metadata stored in both the archival and presentation instances, if one or more parts of the complex architecture should fail, the collection can be restored. Once the publication instance is completed and accessible to the public, users will be able to download selected images and their correspondent DC metadata, containing all the information related to those images.

5. Conclusion

This work was developed for an evolving archaeological dataset, but can act as a model to inform any kind of similarly complex academic research collection. The model illustrates that DC metadata can act as an integrative platform for a non-traditional (but increasingly common) researcher-curated, distributed repository environment. With DC as a bridge between collection instances we ensure that the relationships between objects and their metadata are preserved and that original meaning is not lost. Integration also reduces overhead in entering repetitive information and provides a means for preservation. In the event that a database fails or becomes obsolete, or if ICA can no longer support the presentation instance, the archival instance can be sent to a canonical repository with all its metadata intact.

Finally, we can also attest that the model enables an organized and documented research process in which curators can conduct a variety of tasks including archiving, study, and publication, while simultaneously integrating legacy data. Our whole team, including specialists working remotely, can now access our entire collection as a whole, view everything in context, and work collaboratively in a single place. Because this work was developed with and by the users actively testing it during ongoing study, we can also speak to the real benefits that have been gained. In the course of this work, ICA lost over 2/3 of its research and publication staff due to budget cuts. While this was a serious blow, the collection architecture we have described here has allowed us to radically streamline our study and publication process enough that, despite losing valuable staff, we are actually producing our publications much more efficiently than we ever have before and have helped ensure a future for the data behind them.

References

ADS, Archaeology Data Service. (2014). Retrieved May 9, 2014 from http://archaeologydataservice.ac.uk/.

ARK, the Archaeological Recording Kit. (2014). Retrieved May 9, 2014 from http://ark.lparchaeology.com/.

Cisco, Susan. (2008). Trimming your bucket list. ARMA International's hot topic. Retrieved May 9, 2014 from http://www.emmettleahyaward.org/uploads/Big_Bucket_Theory.pdf.

Corral. (2014). Retrieved August 14, 2014 from https://www.tacc.utexas.edu/resources/corral.

CRM. (2014). CIDOC Conceptual Reference Model. Retrieved May 9, 2014 from

http://www.cidoc-crm.org/.

Eiteljorg, Harrison. (2011). What are our critical data-preservation needs? In: Eric C. Kansa, Sarah Whitcher Kansa, & Ethan Wattrall (eds). Archaeology 2.0: New Approaches to Communication and Collaboration. Cotsen Digital Archaeology series 1, 251–264. Los Angeles: Cotsen Institute of Archaeology Press.

Eve, Stuart, and Guy Hunt. (2008). ARK: A Developmental Framework for Archaeological Recording. In: A. Posluschnya, K. Lambers, & I. Herzong. (eds). Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007. Kolloquien zur Vor- und Frühgeschichte 10. Bonn: Rudolf Habelt GmbH. Retrieved from: http://proceedings.caaconference.org/files/2007/09_Eve_Hunt_CAA2007.pdf.

Faniel, Ixchel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel. (2013). The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. JCDL 2013 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, 295–304. New York: Association for Computing Machinery. doi:10.1145/2467696.2467712

FITS, File Information Tool Set. (2014). Retrieved May 9, 2014 from https://code.google.com/p/fits/.

Harris, Edward C. (1979). Laws of Archaeological Stratigraphy. World Archaeology Vol. 11, No. 1: 111–117.

ICA, Institute of Classical Archaeology. (2014). Retrieved May 9, 2014 from http://www.utexas.edu/research/ica/.

iRODS, A data management software. (2014). Retrieved May 9, 2014 from http://irods.org/.

Kansa, Eric C., Sarah Whitcher Kansa, & Ethan Watrall (eds). (2011). Archaeology 2.0: New Approaches to Communication and Collaboration. Cotsen Digital Archaeology series 1. Los Angeles: Cotsen Institute of Archaeology Press.

LAITS, College of Liberal Arts. (2014). Retrieved May 9, 2014 from http://www.utexas.edu/cola/laits/.

METS, Metadata Encoding & Transmission Standard. (2014). Retrieved May 9, 2014 from http://www.loc.gov/standards/mets/.

OpenContext. (2014). Retrieved May 9, 2014 from http://opencontext.org/.

PREMIS, Preservation Metadata Maintenance Activity. (2014). Retrieved May 9, 2014 from http://www.loc.gov/standards/premis/.

Python, a programming Language. (2014). Retrieved May 9, 2014 from https://www.python.org/.

PHP, A hypertext preprocessor. (2014). Retrieved May 9, 2014 from http://www.php.net.

Rabinowitz, Adam, Jessica Trelogan, and Maria Esteva. (2012). Ensuring a future for the past: long term preservation strategies for digital archaeology data. Presented at Memory of the Worlds in the Digital Age Conference: Digitization and Preservation, UNESCO, September 26–28, 2012, Vancouver, British Columbia, Canada.

Rodeo. (2014). Retrieved August 14, 2014 from https://www.tacc.utexas.edu/resources/data-storage/#rodeo.

TACC, The Texas Advanced Computing Center. (2014). Retrieved May 9, 2014 from https://www.tacc.utexas.edu/.

tDAR, Digital Archaeological Record. (2014). Retrieved May 9, 2014 from http://www.tdar.org/.

VM, Virtual Machine. (2014) Retrieved May 9, 2014 from http://en.wikipedia.org/wiki/Virtual_machine.

Walling, David, and Maria Esteva. (2011). Automating the Extraction of Metadata from Archaeological Data Using iRods Rules. International Journal of Digital Curation Vol. 6, No. 2: 253–264.

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7): e67332. doi:10.1371/journal.pone.0067332

Dublin Core Metadata for Research Data – Lessons Learned in a Real-World Scenario with datorium

Andias Wira Alam
GESIS – Leibniz Institute for the Social Sciences
Germany
andias.wira-alam@gesis.org

Abstract

As a continuation of our work in the datorium project, we provide a service for autonomous documentation and upload of research data. In this paper we discuss and share our experience of developing such a service by using Dublin Core Metadata. Even small and simple, DC Metadata is an appropriate standard to be taken as basic metadata, for instance in the repository systems. The required elements for describing research data are mostly complex, in particular the acquired information about the data, including survey methods, survey periods, or number of variables. DC Metadata cannot cover all elements needed in the research data repository. However, we show that with some extended elements and front-end based manipulations the DC Metadata can be applied usefully in this real-world scenario and support complex description without overcoming the "simplicity" of the standard.

Keywords: research data repository; metadata; DSpace; infrastructure; datorium

1. Introduction

GESIS – Leibniz Institute for the Social Sciences provides services for research in multiple phases of the research process, such as study planning, data collection, data analysis, and archiving and registration. The main targets are data collected from surveys, data from historical social research, as well as scientific publications. Figure 1 shows the research data lifecycle used by the institute to structure its services. Our project datorium belongs to the phase "archiving and registering". We provide a data repository service for social science and economic researchers with a user-friendly tool for the autonomous documentation, upload and publication of their research data, as illustrated in Figure 2. As stated in Linne (2012), the service focuses particularly on small research projects by researchers who do not necessarily work for an institution or are self-funded. A detailed review carried out by the GESIS Data Archive ensures the quality of the submitted data.

Describing research data requires comprehensive and rich metadata elements, such as provided by the Data Documentation Initiative (DDI)[1] or da|ra metadata[2]. DDI can be used not only to describe the research data on study level (general overview of the research data), but also on the variable level - e.g. for information details about variables, questionnaires, and results. The da|ra metadata is now commonly used in assigning persistent identifiers to research data in the context of the DataCite[3] community. Nonetheless, DC elements are the most-used elements for describing resources, particularly scientific resources (cf. Ell et al. 2011; Qin et al. 2013; Malta et al. 2014). Príncipe et al. (2013) also stated that OpenAIRE is starting to move from a publication infrastructure to a more comprehensive infrastructure that covers all types of scientific "output". DC metadata as a fundamental metadata infrastructure for scientific publications is therefore slowly evolving into an infrastructure for research data as well.

[1] See http://www.ddialliance.org/
[2] See http://www.da-ra.de/en/home/
[3] http://www.datacite.org/

The choice of metadata schema or standard is likely not the main focus for researchers looking to publish the data. Researchers need platforms which allow them to publish their data in an easy way, making the data visible and citable (Wira-Alam et al. 2012; Dimitrov et al. 2013). During the requirements analysis for datorium it became apparent that we had to specify the requirements so as to balance simplicity and usefulness. Similar to the lessons learned reported by Wallis et al. (2010), though in a different context, the discussions between the computer and social scientists in the project team started with the question "what should we build for you?" answered by "what could you build for us?".

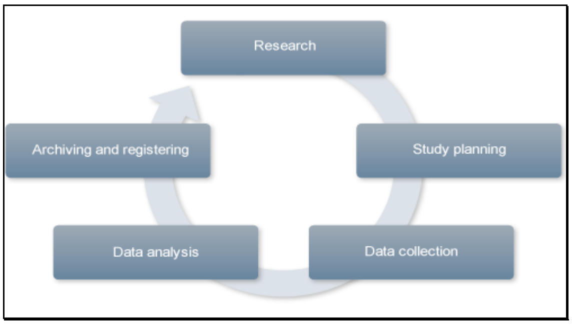


FIG. 1. Research data lifecycle in multiple phases[4]

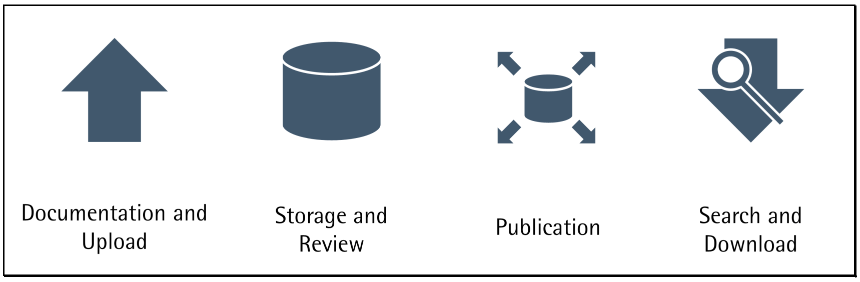


FIG. 2. Illustration of the step-by-step processes within datorium[5]

An ideal vision is that any piece of information in the research data should be well-documented and described. Castro et al. (2013) proposed e.g. to use domain-specific elements in order to fully describe scientific experiments. However, our tool is targeted not only at institution-based researchers but also at any self-funded researchers or even students. Documenting and describing research data is time-consuming and hence expensive work. Thus, increasing the complexity of the documentation process would impact the usability of the tool, and consequently potential users might lose their interest in using it. Accordingly, one of the key challenges is how to make the tool as simple as possible for users, in particular the data depositors, while at the same time gathering as much information about the data as possible. Simultaneously, we have to make the data visible and easy find, especially for the data consumers. Another important feature of the tool is that it shall be available in two languages, namely German and English, in order to target prospective international users: data depositors as well as data consumers.

2. Metadata Design

In the metadata design, we identify not only critical information about metadata in general, but also more detailed information about the research data, e.g. survey / data collection methods, survey periods, or number of variables / units. However, in order to keep the metadata simple, we

[4] http://www.gesis.org/en/services/

[5] https://datorium.gesis.org/

use a Dublin Core subset whose schema is simply flat and has no complex hierarchical structure. As stated by Rice et al. (2008) and Greenberg et al. (2013), datasets are digital materials that need to be described for discovery, preservation, and re-use, e.g. for partner repositories. Furthermore, analogous to the aforementioned work and Greenberg et al. (2009), by using DC elements we provide understandable information about complex objects and help partner repositories or data consumers to become acquainted with the research data. Research data also become more useful when they are interoperable with other data and therefore need a common standard or set of standards (Ball, 2010). As depicted in Figure 3, datorium's metadata schema consists of DC elements and some extended elements. As some elements, e.g. file description, demand hierarchical entries, the schema forms a tree structure. A complex tree structure cannot be described by a schema such as Dublin Core. As mentioned in Chen et al. (2013), research datasets may contain unusual file formats; therefore the uploaded files need additional information e.g. on the number of variables, number of units, languages used in the files, or even software to read the files for further processing. Figure 3 depicts the abstraction of the metadata schema.

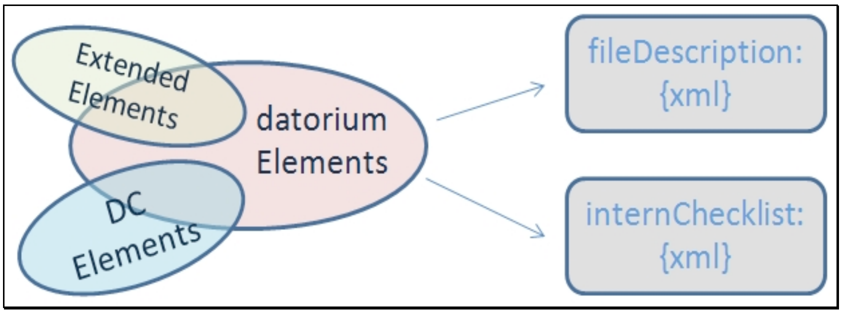


FIG. 3. Illustration of the metadata schema in datorium

To meet the requirement that we need as much information about the data as possible We require 6 mandatory and about 14 optional entries. As mentioned above, we assume that most users are not willing to capture many entries in the tool for simplicity reasons. However we cannot exclude this possibility as there are users who want to provide rich information about their data e.g. to increase the visibility of the data. This situation contrasts with the identified requirements, but we discuss later in the next section how we alleviate this problem. In Table 1 we describe our metadata schema. In comparison with the first design (Wira-Alam et al, 2012), we use 10 DC elements and specify all extended elements with the namespace "dbk" taken from GESIS – Data Catalogue DBK[6]. We also organized the elements in five groups, e.g. General Description or Methodology, according to their affinities. Moreover, we decomposed two elements, Primary Researcher and Contributor, to increase the exactness. In the element Primary Researcher, for instance, we distinguish between person and institution. According to DC standard, however, this property can be filled either with person or institution. Our adaptation makes it possible for users to search only for persons or institutions.

As mentioned above, we support users by increasing the visibility of the data. For this purpose we offer controlled vocabularies, e.g. for Subject Area or Data Collection Method. Controlled vocabularies improve the visibility of the data on the one hand by enhancing the semantic of the metadata, and on the other hand by making the submission process easier for the users. Moreover, in order to support internationalization, we provide all controlled vocabularies in two languages: German and English. This affects both the technical implementation and the search functionality. We demonstrate in the next section how users can benefit from this feature and what the technical implementation looks like.

[6] https://dbk.gesis.org/

TABLE 1: Metadata schema using DC elements and extended elements

	Labels	DC Elements	Extended Elements	Type
General Description	*Title*	dc.title	-	*text*
	Other Title *	-	dbk.othertitle dbk.othertitle.text dbk.othertitle.type (a)	 *text* *text*
	DOI	dc.identifier.uri	-	*URI*
	Primary Researcher *	dc.creator	dbk.primaryresearcher.person dbk.primaryresearcher.institution	*text* *text*
	Publisher	dc.publisher	-	*text*
	Publication Year	-	dbk.publicationyear	*date: YYYY*
	Availability Embargo Embargo (until)	- - -	dbk.availability (b) dbk.embargo.availability dbk.embargo.end	*text* *text* *date: YYYYMMDD*
	Contributor *	dc.contributor	dbk.contributor.person dbk.contributor.institution dbk.contributor.type (c)	*text* *text* *text*
Content	Subject Area *	dc.subject.other	- (d)	*text*
	Topic Classification *	dc.subject.classification	- (e)	*text*
	Abstract	dc.description	-	*text*
Methodology	Geographical Area *	dc.coverage.spatial	- (f)	*text*
	Universe *	-	dbk.universe	*text*
	Selection Method	-	dbk.selectionmethod	*text*
	Data Collection Method *	-	dbk.datacollectionmethod (g)	*text*
	Survey Period *	-	dbk.surveyperiod dbk.surveyperiod.start dbk.surveyperiod.end	 *date* *date*
Additional Notes	Rights *	dc.rights	-	*text*
	Notes *	-	dbk.notes dbk.notes.text dbk.notes.type (h)	 *text* *text*
	Source *	-	dbk.source	*text*
	Publications *	-	dbk.publication dbk.publication.text dbk.publication.id	 *text* *text*
Files	File *	-	dbk.file dbk.file.filename dbk.file.filedescription dbk.file.version dbk.file.versionNumber dbk.file.versionDate dbk.file.resource dbk.file.resourceType dbk.file.resourceTypeGeneral dbk.file.language (i) dbk.file.numberofvariables dbk.file.unit dbk.file.unitNumberOf dbk.file.unitType (j) dbk.file.software dbk.file.alternateId dbk.file.relatedId dbk.file.relatedIdIdentifier dbk.file.relatedIdType (k)	 *text* *text* *text* *text* *date: YYYYMMDD* *text* *text* *text* *text* *int* *text* *int* *text* *text* *text* *text* *text* *text*
Hidden	*Date Issued* *(for sorting purpose)*	dc.issued	- (l)	*date: YYYYMMDD*
	Checklist *(for Curators only)*	-	intern.cheklist	*text*

As explained, the elements consist of DC elements and DBK elements. In the first group, namely General Description, we place all mandatory entries (written in italics). On publication each submission automatically receives a persistent identifier, in this case a DOI® (generated by the system). This increases the visibility of the submitted data by making them as they are citable as scientific publications. We set GESIS – Data Archive as the publisher of the data since they are published through datorium; therefore the value of the element has been fixed and it is not editable. Elements marked with an asterisk are repeatable. We also introduce Other Title in order to accommodate research data that have several titles for some reasons, e.g. original title, translated title in several languages, or project title.

In the second group, namely Content, users can provide a description of the data as a free-text abstract. In addition, users have two important elements: Subject Area and Topic Classification, whose values are controlled vocabularies. The controlled vocabularies provide a possibility for semantic enhancement and thus facilitate connections between the research data and Linked Data on the Web (cf. Isaac et al. 2013). In the group Files we collect relevant information about the files as completely as possible. Further explanation for the elements marked alphabetically from *(a)* to *(l)* is as follows:

(a) dbk.othertitle.type – Type of other title can be selected from the controlled vocabularies provided by DBK (Zenk-Möltgen et al. 2012), such as "project title" or "original title".

(b) dbk.availability – It consists of three controlled vocabularies: "free access", "restricted access", and "embargo".

(c) dbk.contributor.type – Type of contributor is based on the category scheme of the ContributorType from DataCite.

(d) dc.subject.other – Subject Area has been chosen from the disciplines in SSOAR - Social Science Open Access Repository.

(e) dc.subject.classification – Topic Classification is based on DBK[7], consists of overall 38 terms, such as "Economic Systems" or "Social Policy".

(f) dc.coverage.spatial – Geographical Area consists of places, such as countries, cities, or provinces / states, based on ISO-3166.

(g) dbk.datacollectionmethod – Data Collection Method consists of the controlled vocabularies provided by DDI (unreleased beta version, March 2013), such as "Email interview", "recording", or "Telephone interview: CATI".

(h) dbk.notes.type – It is based on DescriptionType provided by DataCite, such as "Abstract" or "TableOfContents".

(i) dbk.file.language – Languages provided by ISO-639.

(j) dbk.file.unitType – Unit Type is based on "Analysis Unit" provided by DDI, such as "Family", "Individual", or "Organization".

(k) dbk.file.relatedIdType – Type of the related identifier is based on the RelationType provided by DataCite, such as "IsCitedBy", "IsDocumentedBy", or "IsPartOf".

(l) dc.issued – Date Issued has been generated by the system at the time of publication.

As mentioned above, datorium offers multi-language support for the controlled vocabularies. The tool supports a so-called ad-hoc translation automatically. Users do not have to take any action in this regard. All controlled vocabularies are stored in a dictionary in two languages. Each vocabulary item in both languages is unique and therefore the correctness of the translation is guaranteed. The controlled vocabularies for Subject Area and Data Collection Method have a tree structure, in opposite of having a long list, to make it easier for users to find and choose the relevant terms for their data.

For the types of the elements, we use rudimentary types for reasons of simplicity. Thus, there are only 4 rudimentary types: *text*, *URI*, *int*, and *date*. Theoretically, with *text* we can cover any types of values. However, we apply a simple validation in order to avoid wrong values. Values typed with *date* without a fixed date format, namely Survey Period, can be given in three

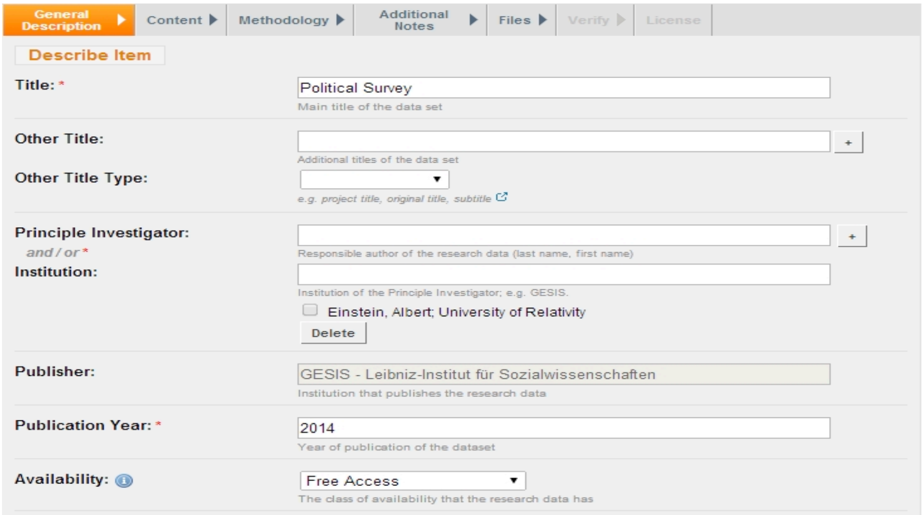
[7] https://dbk.gesis.org/dbksearch/Kategorien.htm

variants: year only (format: YYYY), month and year only (format: YYYYMM), or an exact date (format: YYYMMDD).

3. Technical Implementation

We use a DSpace[8] repository as basis platform for the implementation. The metadata model in DSpace, which is based on Dublin Core, is simply flat and has no complex hierarchical structure. It consists of *schema*, *element*, and *qualifier*. A *schema* is equivalent to namespace, *element* can be considered as content, and *qualifier* can be seen as sub-element if an extra attribute needs to be added. DSpace is a web-based application that follows the Model-view-controller (MVC) architectural pattern (Gamma et al. 1994). This pattern ensures the consistency of the model (data) and the user interface / front-end (view) by employing a controller. DSpace also offers many features such as user management, review process, and discovery / faceted search. Our development process is loosely based on agile software development, which is an iterative process throughout the development cycle.

As described in Table 1, we have groups of elements. In the implementation, we display each group of elements as a tab. This strategy is suitable for data depositors who do not want to spend time capturing information about the data. However, even though all mandatory elements are placed in the first tab, each submission needs to go through all. Figure 4 shows the mandatory and non-mandatory elements in the first tab. For example, the mandatory element Principle Investigator can be filled only by a person, an institution, or both. For the data depositors who are willing to provide as more information about their data, this strategy is also convenient as it provides more structure and orientation for data depositors than a single form with many elements. After the data has been successfully published, the system will assign a persistent identifier (DOI) automatically via a separate module connected with the da|ra API for DOI registration[9].



Title: * Political Survey
Main title of the data set

Other Title:
Additional titles of the data set

Other Title Type:
e.g. project title, original title, subtitle

Principle Investigator: and / or * Institution:
Responsible author of the research data (last name, first name)
Institution of the Principle Investigator; e.g. GESIS.
□ Einstein, Albert; University of Relativity

Publisher: GESIS - Leibniz-Institut für Sozialwissenschaften
Institution that publishes the research data

Publication Year: * 2014
Year of publication of the dataset

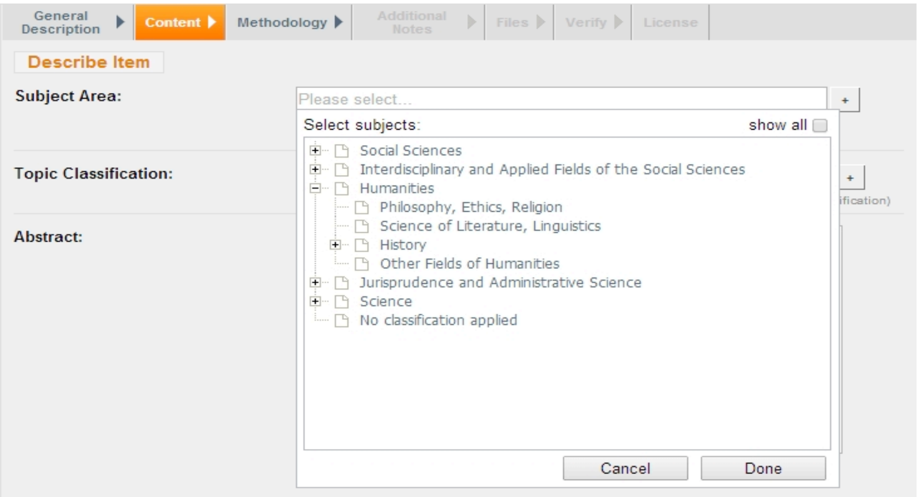
Availability: Free Access
The class of availability that the research data has

FIG. 4. Editor form for General Description

[8] As we mentioned in the previous work (Wira-Alam et al. 2012), we use DSpace (version 1.8.2) as it is an open source repository application. Furthermore, DSpace supports Dublin Core elements by default and has a flat metadata schema which helps us as developers to maintain the data. According to DSpace's website, there are more than 1000+ institutions that have registered to use DSpace for their repository application which is widely used worldwide (May 2014).

[9] http://www.da-ra.de/en/for-data-centers/register-data/

In the Content tab, as captured in Figure 5, one important feature of the tool, namely the controlled vocabularies is shown. We display the vocabularies in their original, i.e. hierarchical form. The hierarchical selection is very comfortable since users can, for example, find or determine an appropriate subject, or more, by its discipline. This feature was implemented without changing the metadata schema. We performed a pure front-end based manipulation and thus the validation occurs in the view as well. A big advantage of this strategy is that the metadata schema becomes flexible since the view does not depend on the model. A possible disadvantage could be wrong values in the database because of a front-based validation level that does not guarantee the consistency. Nevertheless, wrong values only apply for the corresponding element and cannot break the whole elements. Besides, a review process is carried out before the data is published.



Subject Area: Please select...

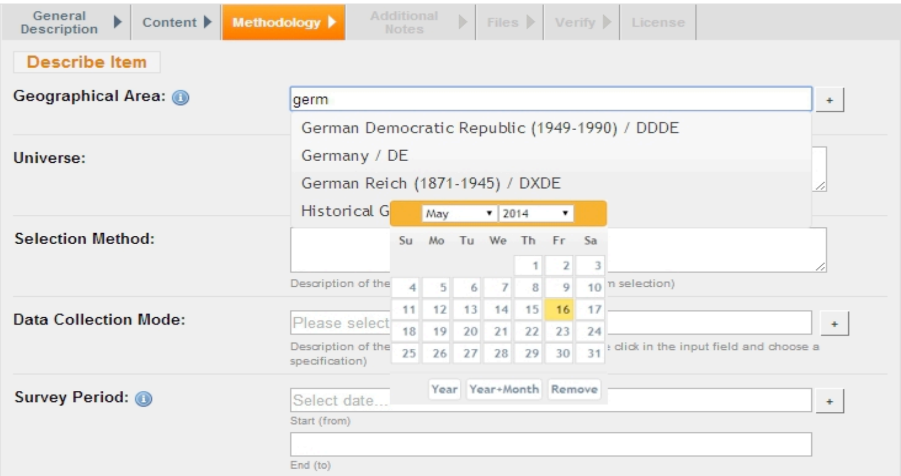
Select subjects:
- Social Sciences
- Interdisciplinary and Applied Fields of the Social Sciences
- Humanities
  - Philosophy, Ethics, Religion
  - Science of Literature, Linguistics
  - History
  - Other Fields of Humanities
- Jurisprudence and Administrative Science
- Science
- No classification applied

Topic Classification: ification)

Abstract:

FIG. 5. Editor form for Content

The next feature regarding the controlled vocabularies is autocomplete. In Figure 6, in the element Geographical Area the data depositors can select places from a given list. Since there are thousands of places to be selected, we provide an autocomplete widget in order to make the selection easier. Users can decide the preferred language (DE/EN); the whole user interface and the controlled vocabularies are then available in the selected language. Another feature is a widget to pick a date. This can be an exact date but also year only or month and year only.



Geographical Area: germ
- German Democratic Republic (1949-1990) / DDDE
- Germany / DE
- German Reich (1871-1945) / DXDE
- Historical G

Universe:

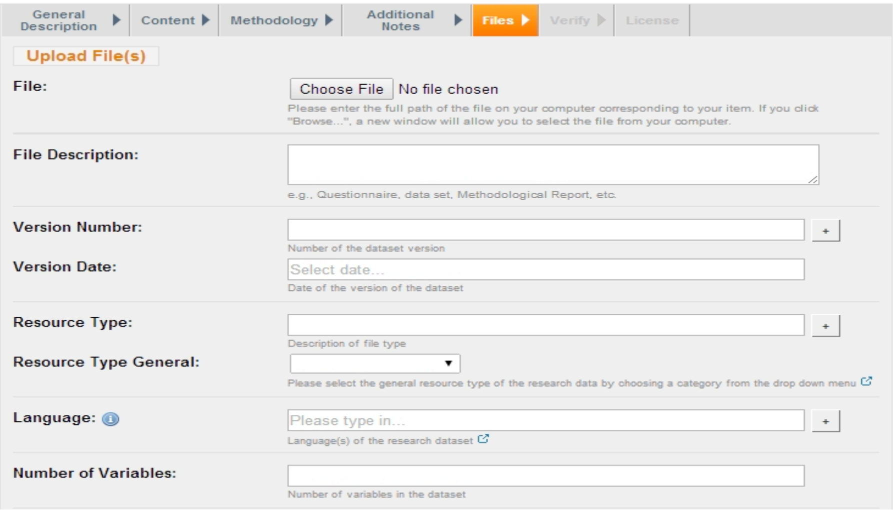
Selection Method:
Description of the n selection)

Data Collection Mode: Please select
Description of the : click in the input field and choose a specification)

Survey Period: Select date...
Start (from)
End (to)

FIG. 6. Editor form for Methodology

As repeatedly mentioned, since we cannot apply a complex metadata schema in the model, we can only modify the view to meet the requirements. For instance, each uploaded file has several elements and each submission / dataset can have several files since it is a repeatable element. This situation therefore leads to a hierarchical form in the model, which is actually not implementable. As shown in Figure 7, we *wrap* these elements in an XML as if they are seen as a single value of the element File to compromise the limitation of the flat metadata model.



File: No file chosen
Please enter the full path of the file on your computer corresponding to your item. If you click "Browse...", a new window will allow you to select the file from your computer.

File Description:
e.g., Questionnaire, data set, Methodological Report, etc.

Version Number:
Number of the dataset version

Version Date: Select date...
Date of the version of the dataset

Resource Type:
Description of file type

Resource Type General:
Please select the general resource type of the research data by choosing a category from the drop down menu

Language: Please type in...
Language(s) of the research dataset

Number of Variables:
Number of variables in the dataset

FIG. 7. Editor form for Files Upload

For the data consumers, finding data is an intellectual effort. In addition to free-text search, faceted search is a well-known technique that helps users to browse large data collections, e.g. images or documents, and delve into more details if required (Yee at al. 2003). By using this technique, and since the controlled vocabularies are available in German and English, the data can be also searched with keywords in a language in which the data was not documented originally. Figure 8 demonstrates the multi-language support for the faceted search. The element Geographical Area shows same values according to the preferred language.

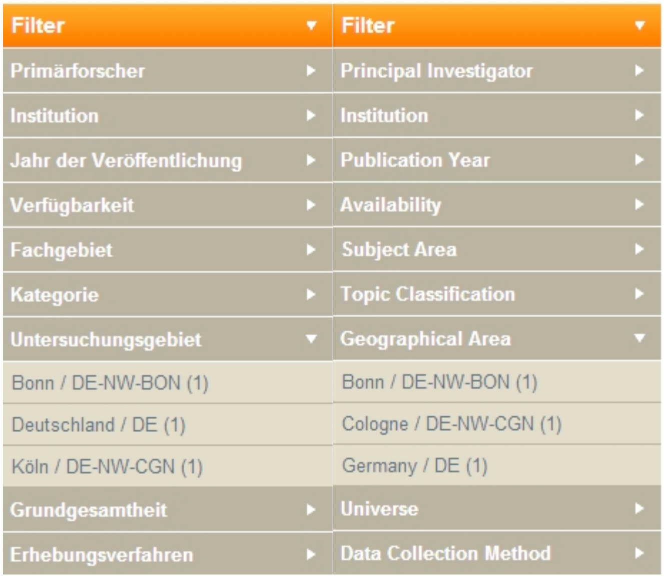


FIG. 8. Multi-language support for faceted search

All front-end based manipulations make use of JavaScript, in particular jQuery[10] and its plugins, and had been successfully tested in various browsers in different versions, among others Internet Explorer, Safari, Opera, Mozilla Firefox, and Google Chrome. The layout and user interface are based on Manakin's XMLUI[11] with many modifications according to the GESIS Web-Style-Guide[12].

4. Conclusion and Future Work

We use Dublin Core for our purpose since it is a simple and appropriate schema for documenting research data. However, to meet all requirements, some extensions are needed. We have shown some approaches to make the application useful and cover complex description without overcoming the "simplicity" of DC metadata. The front-end based manipulation, as we demonstrated in this paper, can remedy the limitation of the schema, e.g. to deal with complex, repeatable elements structures. The documentation of the research data currently refers to the study level; details about the variables used in the survey are not covered. Nevertheless, it is at all times possible to extend the schema so as to meet new requirements. Since the schema and front-end are quite distinct from each other, our approach is suitable for this situation because of its flexibility.

As future work, we want to establish the connection between publication and research data automatically (Boland et al. 2012; Ritze et al. 2013) in order to incorporate scientific publications in research data and the other way around. Moreover, an integrated search with other partner repositories is under way. Therefore we plan to implement an export / import, harvesting (e.g. OAI-PMH) interface, and a schema crosswalk to other standards, e.g. DDI.

Acknowledgements

This project is funded by GESIS – Leibniz Institute for the Social Sciences. We thank our project members, Monika Linne, Wolfgang Zenk-Möltgen, Oliver Hopt, Natascha Schumann, Stefan Müller, Reiner Mauer, and Martin Friedrichs, for their useful inputs on requirements analysis and for their feedback and fruitful discussions during the implementation. Special thanks go to Sigit Nugraha who helped extensively with the technical implementations and Astrid Recker for proof reading this paper. Last but not least, we also thank the DSpace developers for their effort to continually develop this great repository system as an open source project. datorium can be accessed online at https://datorium.gesis.org/.

References

Ball, A. (2010): Review of the State of the Art of the Digital Curation of Research Data (version 1.1). ERIM Project Document erim1rep091103ab11. Bath, UK: University of Bath. Retrieved April 28, 2014 from http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf

Boland, K., Ritze, D., Eckert, K., Mathiak, B. (2012): Identifying References to Datasets in Publications. TPDL, Vol. 7489 of Lecture Notes in Computer Science, page 150-161. http://dx.doi.org/ 10.1007/978-3-642-33290-6_17

Castro, J. A., Ribeiro, C., Rocha da Silva, J. (2013): Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3685

Chen, H., Lin, Y., Chen, C. (2013): Approaches to Building Metadata for Data Curation. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3691

Dimitrov, D., Baran, E., Wegener, D. (2013): Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data. In: Krempels, Karl-Heinz; Stocker, Alexander (Hrsg.): Proceedings of the 9th

[10] http://jquery.com/

[11] https://wiki.duraspace.org/display/DSDOC18/XMLUI+Configuration+and+Customization

[12] http://www.gesis.org/styleguide/

International Conference on Web Information Systems and Technologies: Aachen, Germany, 8 - 10 May 2013: SciTePress, pages 155-159

Ell, B., Vrandečić, D., Simperl, E. (2011): Labels in the Web of Data. In Proceeding of the 10th International Semantic Web Conference, 2011. http://dx.doi.org/10.1007/978-3-642-25073-6_11

Gamma, E., Helm, R., Johnson, R., Vlissides, J. (1994): Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional. ISBN-13: 978-0201633610

Greenberg, J., Swauger, S., Feinstein, E. (2013): Metadata Capital in a Data Repository. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3678

Greenberg, J., White, H. C., Carriera, S., Scherleb, R. (2009): A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata. Volume 9, Issue 3-4, pages 194-212. doi:10.1080/19386380903405090

Isaac, A., Charles, V., Fernie, K., Dallas, C., Gavrilis, D., Angelis, S. (2013): Achieving Interoperability between the CARARE Schema for Monuments and Sites and the Europeana Data Model. Proc. International Conference on Dublin Core and Metadata Applications 2013. http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/171

Linne, M. (2013): Sustainable data preservation using datorium: facilitating the scientific ideal of data sharing in the social sciences. In: Borbinha, José; Nelson, Michael; Knight, Steve (Hrsg.): Proceedings of the 10th International Conference on Preservation of Digital Objects, Lisbon: Biblioteca Nacional de Portugal, S. 150-155. Retrieved May 16, 2014 from http://purl.pt/24107/1/

Malta, M. C., Baptista, A. A. (2014): A panoramic view on metadata application profiles of the last decade. Int. Journal of Metadata Semantic and Ontologies Vol. 9, Issue 1 (February 2014), pages 58-73.

Príncipe, P., Rodrigues, E., Rettberg, N., Schirrwagen, J., Loesch, M., Karstensen, M., Nielsen, L. H. (2013): OpenAIRE Guidelines for Data Archive, Literature Repository and CRIS Managers. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/viewFile/3695

Qin, J., Li, K. (2013): How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/viewFile/3670/1893

Rice, R. (2008). Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008. Retrieved May 16, 2014 from http://dcpapers.dublincore.org/pubs/article/view/945

Ritze, D., Boland, K. (2013): Integration of Research Data and Research Data Links into Library Catalogues. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3683

Wallis, J. C., Mayernik, M. S., Borgman, C. L., Pepe, A. (2010): Digital libraries for scientific data discovery and reuse: from vision to practical reality. Proc. 10th Joint Conference on Digital libraries (JCDL '10). ACM, New York, NY, USA, pages 333-340. doi:10.1145/1816123.1816173

Wira-Alam, A., Dimitrov, D., Zenk-Möltgen, W. (2012): Extending Basic DublinCore Elements for an Open Research Data Archive. Proc. International Conference on Dublin Core and Metadata Applications 2012. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3664/1887

Yee, K.-P., Swearingen, K., Li, K., Hearst, M. (2003): Faceted Metadata for Image Search and Browsing. ACM SIGCHI: Human Factors in Computing Systems, 2003. http://dx.doi.org/10.1145/642611.642681

Zenk-Möltgen, W., Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Reports 2012/1. Retrieved June 21, 2012 from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-292372

Metadata for Research Data: Current Practices and Trends

Sharon Farnel
University of Alberta,
Canada
sharon.farnel@ualberta.ca

Ali Shiri
University of Alberta,
Canada
ali.shiri@ualberta.ca

Abstract

This paper reports a study that examined the metadata standards and formats used by a select number of research data services, namely Datacite, Dataverse Network, Dryad, and FigShare. These services make use of a broad range of metadata practices and elements. The specific objective of the study was to investigate the number and nature of metadata elements, metadata elements specific to research data, compliance with interoperability and preservation standards, the use of controlled vocabularies for subject description and access and the extent of support for unique identifiers as well as the common and different metadata elements across these services. The study found that there was a variety of metadata elements used by the research data services and that the use of controlled vocabularies was common across the services. It was found that preservation and unique identifiers are central components of the studied services. An interesting observation was the extent of research data specific metadata elements, with Dryad making use of a wider range of metadata elements specific to research data than other services.

Keywords: metadata; research data; research data services; standards

1. Data Repositories

"And yet, data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself" (Gold 2007). Although the nature of research data can vary widely depending on the discipline, its importance to the replication, refutation or validation of the findings or observations of a research project has never been in doubt.

Research data has recently been viewed as being part of a larger data landscape, namely big data. A number of researchers have referred to research data, linked data, the web of data and open data as constituting elements of the big data landscape (Hudson, 2012; Shiri, 2013). The *Report of the 2011 Canadian Research Data Summit* (Research Data Strategy Working Group, 2011) provides a specific categorization of digital data, namely research data, produced by academia, industry and government.

The sharing of research data has long been a practice among many research communities, often through informal means made increasingly easy with the advent of the internet and associated tools such as email, ftp sites, etc. Borgman (2007) provides four rationales for the sharing research data, namely "to (a) reproduce or verify research, (b) make results of publicly funded research available to the public, (c) enable others to ask new questions of extant data, and (d) advance the state of research and innovation". She also notes that common metadata formats, ontologies and data structures will support the integration of multiple data sources and services.

The rise of the open data[1] and open science data[2] movements, in conjunction with the increasing implementation of data management and sharing policies by funding bodies[3],

[1] http://en.wikipedia.org/wiki/Open_data

[2] http://en.wikipedia.org/wiki/Open_science_data

[3] http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

governments[4] and journals[5], has led to an explosion in the number of research data services created to serve institutions, association members, and research communities. Databib[6] and re3data.org[7] maintain listings of research data services, and as of August 2014 combined list nearly one thousand. Many services enable the deposit of research data and associated metadata, while others focus on metadata describing research data that is housed in other repositories.

This proliferation of services offering a range of functionalities and designed to serve different communities with different needs poses many challenges to researchers, librarians and others within the research community working to create an interoperable research environment. Documenting the range of functionalities as well as defining means of comparing one service to another have been recognized as important activities and have begun to be addressed by Databib[8] and Dryad[9] respectively. Key to any overall comparison or evaluation is an understanding of the metadata practices within services.

2. Metadata in Data Repositories

Metadata is structured information that provides context for information objects of all kinds, including research data, and in doing so enables the use, preservation, and reuse of those objects. The importance of quality, standards based metadata has long been understood by those in the fields of librarianship and research data management; NISO's six principles of good metadata (NISO 2007) being an excellent and oft-cited expression of that understanding. The same, however, has not always been the case among research communities. A recent study (Tenopir et al., 2011) found that there is a "lack of awareness about the importance of metadata among the scientific community - at least in practice" and recommended that institutions and individuals within them who work with researchers can and should do more to help researchers prepare the metadata necessary to enable the discovery, preservation, and reuse of their data. In a scoping study, Ball (2009) explored the feasibility and desirability of a harmonized application profile to improve resource discovery and reuse of scientific and research data in the repository landscape. The two key findings of his study were that a) a comparison of data models and metadata schemes from a variety of disciplines suggested that a carefully generalized metadata profile could be constructed that is both widely applicable and yet still fulfils the requirements of the use cases and b) while the comparison of several different data models shows sufficient common ground for a relatively detailed data model on which to base a Scientific Data Application Profile, from an implementation perspective a simple model is preferred.

One of the main arguments for the identification and documentation of metadata practices and formats for research data services is to create a solid basis upon which subject and semantic interoperability can be ensured. Identifying useful metadata elements and practices will support various interoperability models reported in the literature (Nicholson and Shiri, 2003; Hafezi, et al., 2010). The same arguments that were made in the first generation of digital libraries, open archives and content management systems hold true for research data services - the variety of disciplines involved and the vastness of research data call for a more systematic and holistic approach to metadata. In their 2012 study, Willis et al. identified 11 fundamental metadata goals for metadata documenting research data and highlighted the need for further metadata-related research. An evidence-based approach to the study of emerging research data management systems allows us not only to study emerging trends but also to develop a basis for formulating

[4] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[5] http://www.plosone.org/static/policies#sharing

[6] http://databib.org/index.php

[7] http://www.re3data.org/

[8] http://goo.gl/mQvy0F

[9] http://www.dcc.ac.uk/webfm_send/750

best practices and policies for research data management. This study aims to take a step towards that goal.

3. Purpose

Given the confluence of increased requirements around data management and sharing with greater demand by researchers for services around metadata standards and applications, an examination and comparison of the metadata standards and practices of research data services would be both timely and beneficial. Given the emerging nature of research data repositories and the urgent need for evidence-based practices, it is important to study examples of the repositories that have been experimenting with how best to organize and manage research data. This is not only useful for the metadata community in conceptualizing metadata standards in a new and emerging context, it is particularly important for planners and practitioners who aim to embark on research data repository projects. The objective of this study is to examine the metadata standards and formats used by a select number of research data services to address several specific research questions. These research questions are concerned with both theoretical as well as practical aspects of organizing, managing and providing access to research data.

1. What is the number and nature of metadata elements available?
2. What research data specific metadata do these services provide in addition to common metadata elements?
3. To what extent do the research data management services adhere to widely recognized interoperability and preservation metadata standards?
4. Which research data repositories benefit from and promote controlled vocabularies for subject description and access?
5. How many of the services provide support for unique identifiers (e.g., DOIs)?
6. What kind of metadata assistance (documentation, etc.) is provided?
7. What metadata elements are common and different across these services?

4. Methodology and Analysis

The nature of this study is exploratory in the sense that it aims to gain an insight into the current metadata practices and trends in four research data services: Datacite,[10] Dataverse Network,[11] Dryad,[12] and FigShare.[13] The rationale for the selection of these services lies in the fact that these are widely popular and internationally used research data services that cover multiple disciplines. A significant number of research-intensive and academic institutions are already using these services and some are considering them in their research data management planning.

Table 1 provides an overview of the geographic distribution of these research data services, their subject areas as well as their main services.

TABLE 1: Research data services

[10] http://www.datacite.org

[11] http://thedata.org/

[12] http://datadryad.org/

[13] http://figshare.com/

Service	Subject area	Main services	Location
Datacite	General	Metadata, DOI	UK
Dataverse Network	General	Cite, analyze, preserve	US
Dryad	General	data underlying scholarly publications discoverable, accessible, understandable, freely reusable, and citable	US
FigShare	General	figures, datasets, media, papers, posters, presentations and filesets, altmetrics	UK

The seven research questions above, which are informed by the NISO principles for good metadata (NISO 2007), provide the analytical framework for examination of research data services focusing on various aspects of metadata elements, formats, and standards. As was stated earlier, an evidence-based approach for this study was thought particularly useful, partly because of the emerging nature of research data management systems and partly because of the variety of disciplines and domains that current research data management services cover. To address the research questions, existing metadata records, metadata creation interfaces, and associated documentation will be examined. The following comparative table addresses the key research questions.

5. Findings

Table 2 provides an overview of our sample set of research data services with respect to research questions 1 through 6.

TABLE 2: Research data services comparison (research questions 1-6)

	Datacite	**Dataverse Network**	**Dryad**	**Figshare**
Number of metadata elements	41	100	52	12
Research specific metadata elements	No	Yes	Yes	No
Compliance with standards	Datacite Metadata Schema, which is an application profile of Dublin Core (DC), OAI	Data Documentation Initiative (DDI) Codebook, compliant with Dublin Core (DC) and Content Standard for Digital Geospatial Metadata (CSDGM), MARC LOCKSS, OAI	Dublin Core, Darwin Core, Bibliographic Ontology, METS/MODS OAI/DC OAI/ORE (Object Reuse and Exchange) RDF/DC CLOCKSS For now, OAI/DC is the recommended format.	CLOCKSS
Use of controlled	Includes controlled vocabularies for	Supports use of controlled vocabularies	Supports use of ontologies and	No formal controlled

vocabularies	some elements, supports use of controlled vocabularies for other elements; MESH, OBI, NCBI		controlled vocabularies such as Open Biomedical Ontologies & Gene Ontology. A trial version of HIVE is provided to support subject description. LCSH, TGN, MESH, Integrated Taxonomic Information Systems (ITIS), National Biological Information Infrastructure Biocomplexity Thesaurus, LC Name Authorities file	vocabularies; only 14 high level categories
Support for DOI	Yes	Yes	Yes	Yes
Metadata assistance	full documentation of metadata schema, user guidelines, full api documentation	metadata documentation available via user guide, contextual help available for each element in metadata entry form	Dryad Wiki pages provide detailed documentation including Cataloguing guidelines	Partner with DataCite

In terms of metadata elements, the services range in number from 12 to 100. Of course, the number of elements is not a measure of success or performance of a system. The number of metadata elements may be dependent on a wide range of factors, including the simple or sophisticated approaches that the research data repositories adopt, the disciplines and domains that they cover as well as the applicability of the elements in terms of metadata creation and maintenance. The proportion of general metadata elements in comparison to research data specific elements ranges quite dramatically; Datacite has no research data specific metadata elements while Dryad has 35 (of 52 total). Dataverse and Dryad provide a more sophisticated set of metadata elements and standards. Figshare takes a minimalist approach and provides a very basic set of metadata elements to facilitate quick and easy deposit of research data.

Preservation appears to be one of the central components of research data services to ensure long term access to data. Most have adopted preservation strategies associated with LOCKSS[14] (Lots of Copies Keep Stuff Safe) and CLOCKSS[15] (Controlled LOCKSS) as widely used and common information and data preservation approaches. Given the importance of interoperability in research data management services, DataCite, Dataverse Network and Dryad support OAI-PMH[16] (Open Archives Initiative/ Protocol for Metadata Harvesting) to ensure the wider findability and discoverability of research data

Initial comparison of several of the sample research data services demonstrates that a variety of metadata standards are in use, although Dublin Core is used or supported across most of the services. Support for controlled vocabularies is common, although few incorporate them by default into their schema. For instance, while Dryad and DataCite adopt a more systematic

[14] http://www.lockss.org/

[15] http://www.clockss.org/clockss/Home

[16] http://www.openarchives.org/pmh/

approach to the use of various controlled vocabularies for subject description and access, recommending various thesauri and knowledge organization systems, Figshare does not provide any specific provision for this feature; the only subject access mechanism in Figshare is the high level subject categories that appear when users click on the 'browse' option on the homepage.

An encouraging sign is the common support for DOIs which are seen as key to discovery, preservation and citation of research data. All of the services appear to have metadata documentation available to aid users.

Table 3 provides a detailed account of the common and unique metadata elements used by the four research data repository services.

TABLE 3: Research data services comparison (research question 7)[17]

	Datacite	**Dataverse Network**	**Dryad**	**Figshare**
Titles	title	- title - subtitle - document title	- article title - journal title - data package title	title
Creators, Contributors	- creator - contributor - publisher	- author - producer - funding agency - distributor - depositor - contact - data collector	- author - creator	- author - collaborators
Topical subject(s)	subject	- keyword - topic classification	- keyword - scientific name	- categories - tags
General description	description	abstract	- article abstract - description	description
Object type(s)	resource type	kind of data	type	type
Date(s)	- date - publication year	- production date - distribution date - deposit date - version date - date of collection-start - date of collection-end	- date of issuance - deposit date - date available - embargo date	- date created - date published
Rights, Access, Use	rights	- data access place - original archive - availability status - confidentiality declaration - special permissions - restrictions - conditions - provenance - document holdings - disclaimer	- rights statement - location of related content outside of Dryad	license
Object technical characteristics	- size - format	- software - software version - size of collection - study completion	- file format - file size - provenance	file size

[17] Note that table 3 does not reference attributes or attribute values and is not meant to be an element by element mapping

Spatial subject(s)	- geo location	- country/nation - geographic coverage - geographic unit - geographic bounding box	- spatial coverage	
Identifiers	- identifier - alternate identifier - related identifier	- study global ID - other ID	- article identifier - associated Dryad data package identifier - data package identifier - identifier for related data in Dryad partner repository - associated Dryad publication record identifier - associated Dryad data file record identifier - data file identifier - issn - electronic issn	
Temporal subject(s)		- time period covered-start - time period covered-end	- temporal coverage	
Citation		- citation requirements - depositor requirements	- journal volume number - journal issue - article start page - article end page - article pages	
Versioning	version	version		
Methodology		- unit of analysis - universe - time method - frequency - sampling procedure - major deviations for sample design - collection mode - type of research instrument - data sources - origin of sources - characteristics of sources noted - documentation and access to sources - characteristics of data collection situation - actions to minimize losses - control operations - weighting - cleaning operations - study level error nores - response rate - estimates of sampling errors - other forms of data appraisal		

Related resources		- series - series information - replication for - related publications - related material - related studies - other references		
Language(s)	language			
Status			- status - article publication status	
Production		- production place		
Additional grant information		- grant number - grant number agency		
Note(s)		notes		

Dryad, Dataverse and DataCite make use of Dublin Core as well as other metadata schemes and standards. It is not surprising to note that there are common metadata elements across these services. Dryad also utilizes Darwin Core, Bibliographic Ontology and its own repository specific elements. While Figshare makes limited use of metadata elements, at least seven out of eleven metadata elements are consistent with Dublin Core. Therefore, one can argue that there is a set of elements across these four services that allow for basic interoperability if a meta-service were to be created for cross-searching and cross-browsing

One of the key questions this study aimed to address was the inclusion or creation of metadata elements specifically for research data. Our comparative analysis of the above research data services shows that there are research data specific metadata elements being used. Dataverse Network and Dryad incorporate metadata elements in this area. For instance, Dataverse makes use of such metadata elements as *date of data collection, data collectors, depositor, deposit date, data specific file types such as raw data, processed data*. Dryad offers a number of metadata elements related to the data package and data files deposited into Dryad. Examples of these elements include: *Associated Dryad Data Package Identifier, Data Package Title, Data Package Identifier, Associated Dryad Data File Record Identifier, Data File Identifier, Deposit Date.*

6. Conclusions and Future Work

This study compared four different research data services in terms of metadata and research data management practices. The results of this study will improve understanding among researchers, librarians and research data managers of the application of metadata in research data services. These preliminary findings contribute to the development of a set of guidelines and best practices for developing and implementing metadata for research data services in order to pave the way for the development of an interoperable research data environment. Furthermore, the identification of metadata elements and formats in commonly used research data services will contribute to the creation of an interoperable research data environment. Future work will include expanding this analysis to additional research data services, both general and domain-focused, as well as comparing in detail the metadata elements common across and unique among the services. The development of a framework that takes into account such important components as preservation infrastructures, unique identifiers, interoperability architecture and the definition of a set of research data specific metadata should guide further research and development in this area.

References

Alipour-Hafezi, Mehdi, Abbas Horri, Ali Shiri, and Amir Ghaebi. (2010). Interoperability Models in Digital Libraries: An Overview. The Electronic Library, 28(3), 438-452.

Ball, A. (2009). Scientific data application profile scoping study report. *June 3rd*. Retrieved August 5, 2014, from http://alexball.me.uk/docs/ball2009sda/.

Borgman, Christine L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Gold, Anna. (2007, September/October). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. D-Lib Magazine. 13/9/10. Retrieved, August 5, 2014 from http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

Hodson, Simon. (2012). JISC and Big Data. Eduserv Symposium 2012: Big Data, Big Deal? May 10, 2012, London, UK.

Nicholson, Dennis and Ali Shiri. (2003). Interoperability in Subject Searching and Browsing. OCLC Systems & Services, 19(2), 58 - 61.

NISO. (2007). A Framework of Guidance for Building Good Digital Collections: Metadata. Retrieved, August 5, 2014, from http://www.niso.org/publications/rp/framework3.pdf.

Research Data Strategy Working Group. (2011). Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit. Retrieved, August 5, 2014, from https://web.archive.org/web/20140312192321/http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/obj/doc/2011_data_summit-sommet_donnees/Data_Summit_Report.pdf.

Shiri, Ali. (2013). Linked Data Meets Big Data: A Knowledge Organization Systems Perspective. Advances in Classification Research Online, North America, 24(1). Retrieved, August 5, 2014, from http://journals.lib.washington.edu/index.php/acro/article/view/14672.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6). Retrieved, August 5, 2014, from http://dx.plos.org/10.1371/journal.pone.0021101.

Wiley, Christie. (2014), Metadata Use in Research Data Management. Bulletin of the Association for Information Science and Technology, 40(6). Retrieved, August 5, 2014 from http://www.asis.org/Bulletin/Aug-14/AugSep14_Wiley.html.

Willis, Craig, Jane Greenberg, and Hollie White. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505 - 1520.

Infrastructure & Models—Part A

The ARK Identifier Scheme: Lessons Learnt at the BnF and Questions Yet Unanswered

Sébastien Peyrard
BnF, France
sebastien.peyrard@bnf.fr

Jean-Philippe Tramoni
BnF, France
jean-philippe.tramoni@bnf.fr

John Kunze
California Digital Library, USA
jak@ucop.edu

Abstract

The Bibliothèque nationale de France (BnF) looks back at lessons learnt over eight years of implementing persistent identifiers (ARKs). While persistent identification is still a relatively young field, this is enough time to gain practical experience, and to conduct a meaningful gap analysis between what is and what should be, especially in a semantic web context. That analysis has exposed important issues concerning best practices and compliance with existing standards.

Keywords: Archival Resource Key; persistent identifiers; web of data; linked data.

Introduction

"Eternity is a very long time, especially towards the end." W. Allen[1]

When considering persistent identifiers, one tends to focus on two ends of the timeline: the immediate near term (at the initial implementation stage) and the very long term, the latter often being too abstract to act on directly. After eight years of implementation experience and almost 20 million ARKs assigned, the BnF now takes the opportunity to look back. This article explores what issues have to be considered during the lifespan of persistent identifiers, in this case ARKs. It also touches on the ARK standard: this 13-year-old standard might benefit from clarification or modification. At a time when institutions are diving into linked data and appear as key stakeholders in the web of data, we believe persistent identifiers have a key role in supporting trustworthy and stable bridges across data silos.

1. The ARK identifier scheme: overview

ARK identifiers have been introduced in various articles and web resources (CDL, 2013) (Kunze, 2003). This section summarizes only enough to make the rest easily understandable.

1.1. Purpose and aim

The ARK standard addresses the same issues as other persistent identification schemes. Although anyone can use them, and there are about 270 organizations currently registered (CDL, 2014), ARKs have been most popular with heritage institutions. These institutions are usually tasked with indefinite retention of content, well beyond expected lifetimes of commercial institutions, and where the perspective is set on the very long term.

ARKs have a very conservative approach to persistent identification. Like URNs and DOIs, ARKs are designed to be independent of DNS and the HTTP protocol; however, they are also designed to work directly in today's web environment URLs, by specifying that the hosting arrangement does not affect identity. For example, these ARKs identify the same resource:

- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m
- http://bnf.example.org/ark:/12148/bpt6k5834013m

[1] http://edition.cnn.com/2006/WORLD/asiapcf/07/04/talkasia.hawking.script

- ark:/12148/bpt6k5834013m

The last of these (with no hostname) is the *core immutable identifier*.

1.2. Anatomy

The base ARK name is typically a completely opaque (meaningless) identifier in order to drastically reduce any pressure to change the identifier string over the long term. For example,

- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m
- This sort of base name is often extended with a qualifier that may be less opaque, as in
- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f19.highres

An actionable ARK (an ARK that works in today's web) has three main parts.

- The *core immutable identifier* itself is mandatory and is designed to be globally unambiguous, persistent and opaque. To that end, it has a structure proceeding from the most general to the most specific (left to right):
  - the *identifier scheme* ("ark:/"), a label that is easy to find by simple text miners;
  - the *Name Assigning Authority* (NAA), which has a 5-to-9 digit NAA Number (NAAN) for opacity. NAAN uniqueness is guaranteed via a registry[2] based at the California Digital Library (CDL);
  - the *ARK name* itself, which should be opaque and is assigned by the NAA; if independent ARK name assignments are performed within a single NAA, the NAA often designates sub-naming authorities corresponding to short prefixes for the ARK name, to ensure ARK names uniqueness.
- The *Name Mapping Authority* (NMA), which enables the identifier to resolve to a resource. The NMA is implemented with a Name Mapping Authority Hostport (NMAH), which in today's web environment is usually an HTTP server. This part can change over the long term, which is why it is optional. Here for example the NMAH is "http://gallica.bnf.fr".
- The optional *qualifier part*, which enables extra services provided by the NMAH using the standard ARK reserved characters "." and "/". At BnF they are often used as follows.
  - Naming sub-parts of a resource (e.g. a specific page in a digitized book). This is achieved by *hierarchy qualifiers* beginning with "/" (/f19 in the example).
  - Naming variants or services of the resource (e.g. a specific version in the lifecycle of a digitized book, or the thumbnail of a given image). This is achieved by *variant qualifiers* beginning with "." (.highres in the example)

1.3. Using ARKs

ARKs raise many of the same issues as other persistent identification schemes.

- **Institutional commitment and policy**. Persistent identification is not a technical problem. It will only work if an institution commits to ensure persistence and global uniqueness over the long term. There needs to be a clearly articulated stewardship policy.
- **Assignment procedures**. Clearly articulated procedures are also required to ensure that assignments are unique and consistently applied to defined resource types. Decisions to be made comprise what ARKs are identifying, which resources are considered to deserve separate ARKs, and which resources should be considered variants of the same ARK.
- **Resolution**. One or more NMAHs are needed to resolve ARKs, each NMA defining a level of service provided with the ARKs. Reliable resolution allows reliable *citation*.

[2] The NAAN registry can be accessed at http://www.cdlib.org/uc3/naan_registry.txt.

ARKs also offer two ways of supporting linked data. Besides using content negotiation, ARK end-users may instead append suffixes, called *inflections*, to gain access to services related to a resource, but without requiring them to remember whole new identifiers. For example,

- http://texashistory.unt.edu/ark:/67531/metapth346793/ *(ARK for the resource)*
- http://texashistory.unt.edu/ark:/67531/metapth346793/? *(its metadata)*
- http://texashistory.unt.edu/ark:/67531/metapth346793/?? *(the NMA's commitment)*

By itself an ARK should lead to the resource (object). Appending a single "?" should lead to the resource's metadata (Kunze, 2010) and appending "??" should lead to metadata describing the kind of persistence to expect. In the current archival environment, the latter is critical for indicating when a resource is truly invariant, or subject to correction, or is a growing resource. As an alternative to content negotiation, ARK inflections are easier to use and more precise. Inflections are not as easy to support, however, with the Tomcat-based web services at BnF.

2. A brief history of ARKs at the BnF

2.1. Adoption and initial implementation of the ARK identifier scheme

In 2006, the BnF conducted a risk-driven requirements analysis to adopt the ARK persistent identification scheme. Two core requirements used for selection criteria were (1) *financial independence* of the NAA: identifiers subject to a fee, such as DOIs, were discarded and (2) *technical independence* of the naming authority (since identifiers had to be directly integrated into our in-house Information Systems): identifiers relying on installing special-purpose software, such as Handles, or on external services, such as PURLs, were discarded. BnF needed *stable, location-independent URLs*, which do not redirect to temporary URLs (avoiding the overhead of managing an endlessly increasing number of redirects).

URNs also fit our criteria fairly well, but the ARK specification addressed some areas more precisely than URNs, such as the definition of a persistence policy, and additional services on a particular resource in a web context (through the use of qualifiers). Like the URN scheme, the ARK scheme does not mandate use of one particular vendor or service for its identifiers. Unlike URNs, DOIs, and Handles, however, ARKs also do not mandate use of one well-known DNS resolution starting point, so ARKs can be implemented directly on a local web server. While some consider this a weakness, citing the "inherent" fragility of DNS names, their argument usually suggests using dx.doi.org, handle.net, or n2t.net instead; the logical flaw is that these are DNS names too, and we note that none of them are as long-lived as bnf.fr. The bottom line is that ARKs are implementable with the simplest of technologies, and they do not require a special-purpose global infrastructure uniquely built for their own scheme.

At this stage, ARKs were defined for two distinct types of resources: **digitized documents**, available in the digital library Gallica – using http://gallica.bnf.fr as NMAH and **catalogue records**, which needed identification for exchange with BnF's OAI repositories – using http://catalogue.bnf.fr as NMAH.

For both NMAHs, we defined an initial complete set of qualifiers to name subparts and variants. As an illustration, in gallica.bnf.fr, we defined qualifiers to name the pages of a book (e.g. http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f10 to name page number 10 in the digitized document, /f10n5 to name the set of pages 10 to 14), and qualifiers to invoke variants of a book or a page (http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f10.highres, .medres, .lowres and .thumbnail for the different resolutions of the same page; .text to access the OCR for a particular page, .vocal to access the sound version of the same page). For the main catalogue, qualifiers were used to name distinct formats of the same record.

More details about the initial approach and the first implementation choices are available in (Bermès, 2006). During the eight following years, ARKs became the lingua franca across the institution, and their use expanded to new areas.

2.2. Fostering ARK identifiers: new resources, new clients

Since 2006 BnF has expanded its initial use of ARKs for two different purposes:

- Identifying descriptive records in order to manage them in our **OAI** repositories, and more recently, in our data.bnf.fr **linked data** services. This led to assigning ARKs to EAD finding aids, manuscript illumination records, museographic descriptions.
- Preserving digital documents. In 2010 our preservation repository, SPAR (Scalable Preservation and Archiving Repository), went operational. As each Information Package had to have a persistent identifier, SPAR played the role of an ARK assigner whenever there was no pre-existing ARK assigned to the ingested document.

These different resource sets had different scales and creation workflows, which made it very difficult to have a single ARK assignment procedure. The most central assigner is SPAR, but it is only for digital documents (not descriptive records) and it was rolled out after the assignment channels for mass digitization were operational and optimized, which led to path dependence. On the descriptive records end, some databases had much smaller datasets than the 15 million records of the catalogue, which made semi-automated assignment procedures more suitable.

In the end, ARKs were assigned using three different means.

- **Automated, based on an existing number**: used for our two legacy systems (Gallica and our catalogue records), and for our finding aids database. Our large datasets have pre-existing reliable numeric ids that we can "dress" as ARKs. E.g. the record n°32915216 from the main catalogue had the "c" sub-naming authority for descriptive resources, and the "b" 2nd level sub-naming authority for records from the main catalogue. Thus, 32915216 became ark:/12148/cb32915216j (with the addition of a final check character).
- **Automated, independent of any number**: used for medium to large datasets with no reusable id (because significant or incompatible with the ARK structure). Our preservation repository, SPAR, automatically assigns an ARK upon ingest. E.g. ark:/12148/bc6p01zndd assigned to a web archiving container file, indicates (to BnF staff) that assignment was routed to sub-naming authority "b" (digital content) and to repository "c6p0", a 2nd–level sub-naming authority that takes care of uniqueness at repository level.
- **Semi-automated**: with a list of ARKs that curators assign to resources (one spreadsheet per sub-naming authority), this is used for very small datasets. It meant defining a sub-naming authority per database to guarantee uniqueness. E.g. ark:/12148/cdt9x5ww identifies a book binding description. As a descriptive record, assignment was routed to sub-naming authority "c", then to 2nd-level sub-naming authority "dt9x" for book binding.

On the access side, as new services were being built upon new resources, several ARK NMAHs could be used simultaneously for the same resource[3]. For instance, the same catalogue record can be displayed in the main catalogue, which delivers a "full" but isolated record in traditional formats, and also in data.bnf.fr, which provides the RDF view of this record, but displays it in an enriched landing page that aggregates related resources. The difference is obvious for authority records, which can be seen, for instance, between these two ARKs:

http://catalogue.bnf.fr/ark:/12148/cb118905823

http://data.bnf.fr/ark:/12148/cb118905823[4]

[3] This might be considered a risky practice, as with several NMAHs for the same ARK identifier, you need to know all the NMAHs of a particular resource to have a complete view of it. We addressed this problem by defining a default NMAH for a given resource that is considered the "master" view for such a resource. For instance, http://catalogue.bnf.fr (main catalogue) is the default for bibliographic descriptions. A strength of potentially distinct NMAHs for a single ARK is that it forces one to dissociate the resource from the current application providing access to it, which forces one to adopt a long term perspective.

[4] As of 2014, July, data.bnf.fr accounts for only 60% of the catalogue data. Therefore, 40% of the ARKs in the main catalogue are not (yet) in data.bnf.fr.

2.3. At international scale: backing up the ARK registry

The NAAN registry maintained by the CDL described in §1.2 is a cornerstone for the viability of ARKs, because the centralized registration of NAAs ensures the uniqueness of each NAA Number (NAAN). To this end, it was important to guarantee its persistence over the long term, which led to registry mirroring arrangements with the US National Library of Medicine (NLM) and the BnF. From the BnF point of view, it meant formalizing a partnership with the CDL with a Memorandum of Understanding. As this MoU had to be signed off by the president of the BnF, it had the beneficial side-effect of securing institutional commitment for ARK identifiers from top-level management.

3. Implementation gap analysis: Consolidating ARK curation at the BnF

The previous section describes how ARKs gained momentum at the BnF and were progressively applied for different purposes and resource types beyond the originally envisioned use cases. This led to a wide variety of implementation choices and management rules, and consequently a call for centralized policy and harmonization. A gap analysis was conducted in 2014 to address this question in a systematic fashion. It consisted of summarizing the lessons learnt and problems encountered over the past 8 years, and then organizing those lessons around the following focal areas: functional, organizational or technical issues, qualifier implementation questions, policy descriptions, and compliance with standards. Those focal areas are described in separate subparts of §3 and §4.

The next subsection summarizes the issues uncovered by the gap analysis. Most of them are not complicated technical issues, but rather simple observations that we think would likely be made by any organization similar to BnF after 8 years of managing persistent identifiers.

3.1. Organizational issues

A persistent identifier and its policy should outlive its initial implementers. Obvious as this statement sounds, its direct implications are not readily apparent in the early implementation stages. It requires continuous improvement and refinement of the identifier policy and uses, which must remain stable while accommodating new and evolving uses and needs. This prevents identifiers from falling into obsolescence or disgrace, with a decrease in perceived relevance or visibility. Neither must they become "over-used"; frequent or casual assignment leads to misuse. A disciplined approach to organization and communication are key factors to sustainability.

In eight years, there has been a good deal of staff turnover in the ARK BnF expert team. Only one person from the original seven-member team remains. What's more, as ARK use expanded to new areas (as addressed in §2), its audience got much wider than the original team. This includes **library curators** that use, or might use, ARKs to cite resources; **digital object curators**, that handle the lifecycle of the object, including identification and access; **web application managers**, on the IT and librarian sides; **linked data experts**, especially for the data.bnf.fr project. As a result the communication and documentation had to be adapted for the larger audience, which needed to be aware of policy and key curation issues without necessarily understanding all the details.

Our "ARK consolidation approach" had two organizational phases.

Communicate: gather all the users, train them in the main underlying concepts of persistent identifiers, common misconceptions about them and best practices, and mandate two "reference ARK coordinators" – one on the IT and one on the librarian side.

Set up targeted working groups, led by the "ARK coordinators", these focused on specific resource types or applications, reducing the identified gaps and addressing new needs.

3.2. Functional gap analysis

The functional gap analysis itself revealed many areas for improvement in our persistent identifier services, particularly for resolution and associated services[5].

- Some applications do not create resolvable ARKs, but only record them as metadata.
- Whenever a resource is not available in the ARK-aware URL, there is only a 404 or 403 browser response, which should be replaced by one of the following more explicit statements: 1) *Resource not found* – this is an incorrect URL and no resource has ever been available at this URL; 2) *Resource deleted* – the resource was there, but it was deleted; in this case, provide core metadata and if possible the reason for the deletion; 3) *Access disallowed* in this context; as with deletion, one should provide core metadata and if possible the reason of the withdrawal (e.g. copyright status).
- Across some applications there are obsolete or inconsistent ARK redirects. E.g. an old test version of the digital library, gallica2.bnf.fr, no longer redirects to gallica.bnf.fr.

In all these cases, our minimum baseline service is clearly not achieved. Our first goal is therefore simple but attainable: define BnF "ARK core services" that any persistent-id aware application should comply with, namely,

- Provide access to the object behind the ARK
- In case of object unavailability, provide metadata to understand what was there and why access is no longer possible.
- Set up a generic process for updating redirects at the level of the BnF "ARK coordinators".

3.3. Refining the identification and persistence policies

When ARKs were first implemented, we had an unclear view of what stewardship promise we could return with identifiers. Therefore we ended up with a very high-level statement[6]:

- No identifier re-assignment;
- Identifier string policy: opaque strings, no vowels, use of a final check character;
- Persistence policy: guaranteed, but needs to be refined in the future; the form of the underlying resource can change to ensure its persistence (e.g. format migration).

With almost a decade of experience managing ARK identifiers, digital preservation objects (PREMIS Maintenance Activity, 2012), and alignments between our catalogue records and other linked data sources, we can see possibilities for differentiated persistence policies.

- For a digital document that we preserve, our aim is to keep the information content stable and accessible and useable to end-users. This means permanent access with stable content.
- For a catalogue record, the information content can be updated as the catalogue record is corrected, enriched, updated, etc. This means permanent access with somewhat more dynamic content.
- For an archival records document, the identifier will be maintained but the content may be suppressed for legal reasons. In this case, we provide a "tombstone" with the metadata and reasons for the object unavailability.

The BnF is currently considering formalizing these policies in a systematic way.

[5] This analysis is limited to ARK implementation. Time permitting, BnF could have studied additional identifier systems to get ideas for improvement, however the ARK scheme, being built upon experience with other schemes, was already a leading choice, so a broader study was not considered a priority.

[6] http://gallica.bnf.fr/ark:/12148/btv1b8451622d.policy

3.4. Refining the qualifier implementation

One issue we have to deal with is proliferation of identifier qualifiers (introduced in §1.2), in response to which we decided to create a consistent qualifier policy. From the most generic service to the most specific, we see three tiers of qualifiers.

- *Generic qualifiers.* Applicable to any resource, these are qualifiers providing a description of the resource (.description), its persistence policy (.policy), and potentially a qualifier revealing the sub-parts and variants available for the object.
- *Content-type-dependent qualifiers.* For digitized documents, you can use generic display resolution variants (thumbnails, low, medium or high resolution). For descriptive records, you can use generic metadata formats (RDF, XML…). The list of possible qualifiers can be maintained independently of any application.
- *Application-specific qualifiers.* These are specific to a particular NMAH.

We also consolidated our policy about when it is appropriate to define a new qualifier, due to two considerations revealed in the gap analysis. The first has to do with *querying vs. citations*. Variant qualifiers are *not* a query language, but do allow citation of services that one considers "persistent" and relevant from an end-user point of view. In that light, http://gallica.bnf.fr/ark:/12148/bpt6k65581775.r=food, in particular, the ".r=" qualifier raises a red flag. This qualifier can be viewed as a way to search for a word in a digitized document; but ARK qualifiers are intended to refer to the document, not to "look" into documents. It can also be viewed as a way to act upon a document by returning it (from BnF) "with highlights" added (here on the word "food"). This use case could comply with ARK qualifiers, but the side-effects could be distracting if not misleading. Unfortunately, it is easy to do accidentally; if a user previously searched for a word in a document before copying and pasting the URL, it will include the "r=word" qualifier. In the end, this creates a reference to a document with highlights, whilst most of the time all the user wants to do is refer to the document without them. This means that, in most cases, revealing such parameters is *not* recommended for persistent URLs.

A second consideration is *technical vs. non-technical* qualifiers. Any qualifier that concerns a detail of implementation, technology, or a temporary information object should not be expressed in the URI. Unlike the "ARK name" part, qualifiers are not meant to be long-term persistent. However, their stability and maintenance is important for the perceived trustworthiness of the service, and it is costly. Supporting the aforementioned .r= qualifier has a cost, as the syntax for searching for several words ".r=word1+word2+wordn" has to be maintained over re-implementations.

As a result of this gap analysis, the BnF intends to raise awareness of good practices among ARK users (developers and web application managers) and to formalize a general best practices document. A list of qualifiers will be created and maintained for the three aforementioned levels.

3.5. Technical issues: consolidating the technical framework

From its first implementation, the ARK resolver at the BnF had to meet two basic requirements: complying with the security policy of the IT operations service and managing the increasing flow of network requests.

Initially, the ARK resolver was a part of a general-purpose document viewer application. For each domain-specific application, every incoming URI including an /ark:/ pattern had to be detected by an HTTP reverse proxy and redirected toward this viewer application. The ARK resolver had to analyze the ARK identifier and the request, change it to a domain-specific format, and then forward the request for processing to the domain-specific application. These applications were hosted on multiple servers using virtual IP and load-balancing in order to share the load between these servers. This architecture had some shortcomings. First, the use of a reverse proxy conflicted with the IT operations requirements. Second, to detect, change and redirect the requests, the ARK resolver had to implement some domain-specific rules. This was dangerous for the security and maintainability of the whole system.

After this first architecture was in operation for two years, it was agreed to define a new system that would be more generic, parameterized, and scalable. The multi-server load-balancing system was kept, but three modules were added.

a) A domain-specific module that checks if the incoming request is in the scope of the domain, and if not, sends it to the *ARK redirection module*. This filter module is generic but uses domain-specific patterns to verify incoming requests before they go to module b).

b) Domain-specific sub-modules analyze the request, and if necessary, reformat it according to the domain's requirements before transferring it to the domain-specific application.

c) The *ARK redirection module* is able to analyze the ARK identifier and the incoming request and then forward the request for processing to the domain-specific application. The redirection rules are parameters defined in an XML file.

The new document viewer application is now leaner because it does not handle the resolution of ARK requests. This task has been distributed between the generic redirection filter, the specific reformatting filters, and the centralized ARK redirection module. The workload of this redirection module is lower since many of the incoming requests are going directly to a domain-specific application that can resolve the ARK identifier.

Three years later, new requirements came out in parallel with new developments of the Gallica viewer module. Some tools were implemented to manage ARK identifiers and qualifiers, which are now defined by a configuration file. The processing of ARK qualifiers gained leverage by becoming more generic, which made them easier to use in the Gallica API. The ARK redirection module was enhanced by migrating the old redirection rules to mapping tables stored in a database. That module is also using a copy of the ARK NAAN registry that is mirrored at regular intervals from the NAAN registry at the CDL. The new architecture is summarized in Figure 1.

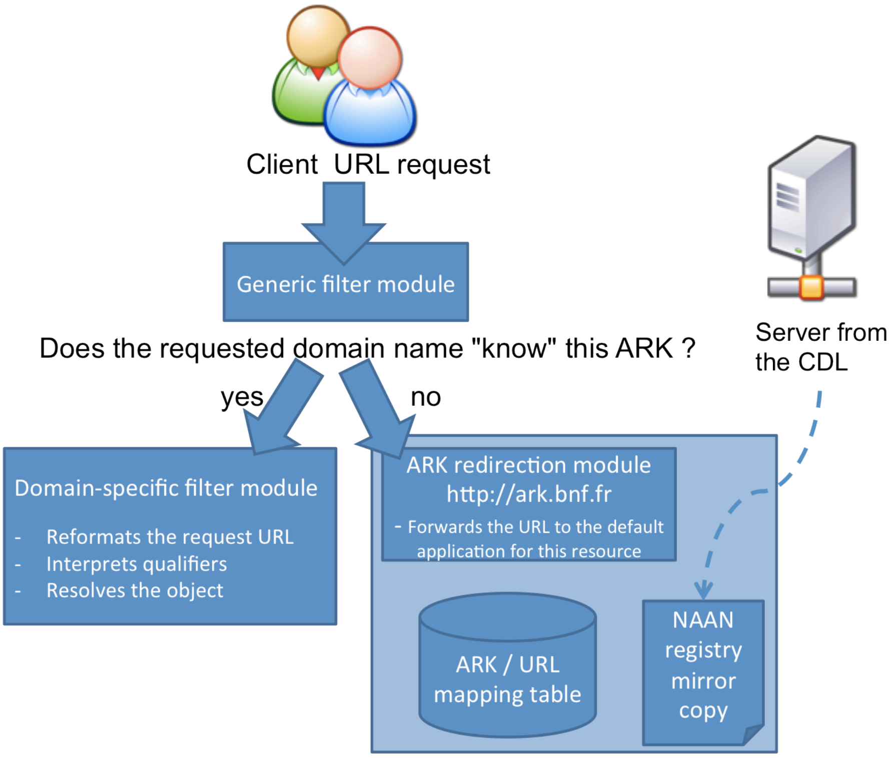

Client URL request
Generic filter module
Does the requested domain name "know" this ARK ?
yes
no
Domain-specific filter module
- Reformats the request URL
- Interprets qualifiers
- Resolves the object
ARK redirection module
http://ark.bnf.fr
- Forwards the URL to the default application for this resource
ARK / URL mapping table
NAAN registry mirror copy
Server from the CDL

FIG. 1. BnF ARK resolver architecture

The ARK minting process, functional aspects of which are outlined in §2.2, has followed a similar evolution. Initially, ARK creation was completely delegated to domain-specific applications. This method was easy to implement but problematic in terms of maintenance and robustness. With implementation of the SPAR repository came the development of a generic function to mint new ARKs. A growing proportion of new identifier assignment is now performed by this generic function.

Since its early stages, the ARK system at the BnF has been tuned regularly to become easier to maintain and configure, although technical issues still remain. To keep a robust system that can be trusted by end-users, we have to consider an increasing diversity of applications, the number of ARKs involved, and the flow of incoming requests.

3.6. Main lessons learnt about persistent identifier curation

To allow operational persistent identifier curation at a non-expert level, core questions have to be answered. With our eight-year hindsight, the key questions could boil down to this check-list:

- Who should be contacted in your institution when new kinds of objects are to be given persistent identifiers or when persistent-id aware applications are defined or revised?
- What are your identifiers identifying?
- Will your identifiers be re-assigned over the long-term or not?
- How much can the underlying content change over time? Can objects be deleted?
- Which services and subparts do you want to reveal, if any, so that end-users can cite a specific portion of the resource and/or a particular variant of that object?

4. Standards gap analysis

4.1. Machine-readable commitments

No identifier, regardless of scheme, can tell us if it will prove to be persistent into the future. The best "it" can do is to tell us (via its NMA) enough about itself, its resource, and resource provider to help us judge how and when to use it. The story it tells must be able to convey such things as provider support policies, expected changes to the resource (e.g., none, or corrections only), and the nature of the provider itself. A persistence promise is not black or white. Instead it is multi-dimensional, suggesting a breakdown into metadata elements.

Because we assume people searching for resources at scale will usefully want to filter based on persistence promise attributes, it will be necessary to support machine-readable commitments expressible via metadata. As was described earlier, the ARK inflection, "??", is designed to gain access to metadata statements about providers' persistence promises. Unfortunately, the ARK standard does not specify how to create machine-readable persistence promises. This section explores some of the areas that metadata should cover in such machine-readable commitments.

Support policies

Support policies and commitments vary between institutions, collections, and even between resources within a collection. For example, users often expect unchanging content behind durable links to *published* content, but they expect dynamic content behind durable links (persistent identifiers) to *advertised* content, such as a home page, curated database, or per-second updated stream of sensor data.

Setting expectations about this "content invariance" (or lack thereof), is critical, because audiences often avoid one kind and seek out the other kind, or vice versa, depending on the situation. Both are legitimate uses of persistent identifiers. Prior work at NLM (Byrnes, 2000) suggests at least four kinds of content invariance:

- *correctable*: Previously recorded content may be corrected (only) at any time.

- *dynamic*: Previously recorded content may be overwritten arbitrarily at any time, provided the resulting new content continues to match its metadata description. For example, the NLM homepage and the local weather page may both advertise very persistent identifiers for content that is completely overwritten from time to time.
- *unchanging*: Previously recorded content will not change, but encodings and markup may change during a format migration.
- *bitstream*: The bitstream representing previously recorded content will not change.

Datasets that grow

There is an important dimension of content invariance describing resources that grow, but whose growth pattern does not alter previously recorded content. We might describe such resources as subject to *non-disruptive growth,* as it is concerned with growth that does not in itself disrupt or displace previously recorded content. This applies to many common information resources, such as live, sensor-based data feeds, citation databases, and even serial publications.

The nature of the provider

Anyone can promise anything, but we might value a promise from one source more than from another. Relevant factors include not only what a provider promises in regard to identifier and resource support, but also how that provider is motivated, supported, and perceived. Thus mission, profit motive, succession plan, and reputation come to bear. Work to be done includes expressing these via metadata.

Support level

What are the provider's naming practices? How often is the collection inspected for broken identifiers? What action is taken when outages occur, and at what priority? Realistically, not all resources are equally important to a provider and its audience. To better support some resources means lowering priority support for other resources. What is a resource's "track record" and can one inspect it? These are all questions that can inform user choices of identifier.

4.2. Using ARKs in a semantic web context: investigating best practices

When the ARK specification came out in 2001, the core semantic web concepts and standards were already out or on their way (RDF was released in 1999). However, as the semantic web gained wider adoption, new best practices about URIs emerged over the next decade (W3C, 2008) and it is timely to re-evaluate the ARK specification in this new context. The main observation is that on one hand, ARKs can be embedded in URIs, which allows their use in the web of data, but on the other hand, the linked data best practices call for "Cool URIs" that, among other properties, "don't change" (Berners-Lee, 1998). For institutions that implement them, ARKs are a natural way to push identified resources onto the web of data. The question now is how to reconcile these two normative contexts at the BnF while implementing ARKs on the data.bnf.fr linked data service.

One could first ask how those two contexts address the question of multiple representations of a resource. On the semantic web, content negotiation using a generic URI yields the relevant representation of a resource; whether to reveal specific URIs for the variants is up to the content provider to decide. There is no reason why a provider could not implement an unqualified ARK name and rely on content negotiation to return linguistic or format variants to the user; or the user can reveal these variants by using traditional qualifiers[7].

[7] For the moment however, data.bnf.fr does not use ARK-URIs for its content negotiation. Early in the project when such choices were made, non-opaque URIs were considered better for SEO, as visibility on the web was one of the core aims of data.bnf.fr. Therefore, http://data.bnf.fr/ark:/12148/cb118905823 redirects to the temporary URI http://data.bnf.fr/11890582/charles_baudelaire/, which provides access to a particular representation of the object depending on the result of content negotiation (RDF/XML,

However, the real question is about the form of the URIs. In the early semantic web, a good deal of debate was about "real-world resources" that can be described on the web of data (with URIs), but could only be put *on* the web via substitutes (e.g. a description and/or a web page). It was initially considered wiser to use non-dereferenceable URIs. Non-HTTP URI schemes like "urn:" could be used to that end, and "info:" was explicitly defined for that purpose. By the end of the 2000's however, there was global consensus that an HTTP URI could be used for any resource. As a result, putting resources on the web of data now implies using HTTP URIs, i.e. URLs. This poses no conflict with ARKs since they are designed to be embedded in URLs using an NMAH that resolves them.

The main conflict between ARKs and URIs used on the semantic web concerns the qualifier part. At issue is distinguishing between a descriptive resource (available on a web page) and its underlying content (which might, or might not, be interpreted as a web page):

> *"*It is important to understand that using URIs, it is possible to identify both a thing (which may exist outside of the Web) and a Web document describing the thing. For example the person Alice is described on her homepage. Bob may not like the look of the homepage, but fancy the person Alice. So two URIs are needed, one for Alice, one for the homepage or a RDF document describing Alice. The question is where to draw the line between the case where either is possible and the case where only descriptions are available." (W3C, 2008).

With ARKs, the URI to reference the descriptive resource is constructed by adding the "?" inflection to the URI of the content resource. Unfortunately, supporting the single "?" (what looks like an empty query string) directly was impossible with the BnF infrastructure. What's more, BnF made the implementation choice to create ARKs directly for descriptive resource (e.g. authority records), so the mechanism needed was the opposite: *from* the identified descriptive resource (identified with an ARK name) *to* its underlying content resource, not the other way round. Therefore, we had to consider the other two mainstream choices:

- "**suffix hash URI**": you have http://example.com/resource for a web resource (e.g. a web page about a person), and http://example.com/resource#classifier for the underlying thing (e.g. the person itself). A browser client automatically strips off the # for consumption, which relies on standard web architecture and best practices.
- "**prefix slash URI**": you have http://example.com/doc/resource for the web document and http://example.com/id/resource for the underlying thing. This requires an HTTP 303 redirect from the resource URI to the URI of the web document.

The semantic web best practices highlight an area currently unaddressed by ARK qualifiers: how to name the underlying "thing" when the ARK is assigned to a descriptive resource. This is clearly not a whole-part problem (addressed by "/). Neither is it really a "service" or "variant" qualifier (addressed by ".") because the two identified things are quite distinct.

With ARKs only the "prefix slash URI" strategy is possible for the current state of the standard, which means using e.g. http://data.bnf.fr/**id**/ark:/12148/ark:/12148/cb118905823 (the French poet Charles Baudelaire) and http://data.bnf.fr/**doc**/ark:/12148/cb118905823 (the record describing him). This was not implemented because the redirection rules would present too great an extra server burden for our application.

From a technical standpoint, in data.bnf.fr the decision was made to locally extend ARKs and use "hash URIs". For example, we separate http://data.bnf.fr/ark:/12148/cb118905823 (web page about Charles Baudelaire) from http://data.bnf.fr/doc/ark:/12148/cb118905823#foaf:Person (Charles Baudelaire himself).

Notation3, N-Triples, JSON, or HTML, and language variants). We intend to reconsider this question with the evolution of SEO practices.

Looking back at the standard, would accommodating this change mean defining a new kind of qualifier, beginning with #, to name the underlying resource? Though technically possible, this would cause backwards compatibility issues, because the # character is not reserved in ARK names. In other terms, one could perfectly define the following (unqualified) ARK core identifier: ark:/9999/c5j3r4#hz45, with a # in the ARK name itself. Defining a # qualifier would break backwards compatibility in such cases. On the other hand, # already has a use in the standard web architecture (fragment for a URL) which makes it unlikely that implementers will use this character in their own implementation. A comprehensive survey of ARK implementers would be useful before any decision. If a # qualifier proved to be possible, we believe this would be a valid scenario to reconcile semantic web and ARK implementation approaches.

Conclusion

This article intended to look back at the history of using ARK persistent identifiers in one institution, and possible evolutions of the standard. Standards-wise, the question boils down to whether we should consider expanding the core features to increase cross-resolver interoperability and adapt ARKs to new contexts, or should we stick to the current ARK recommendation, which is flexible, simple, easy to use, and in most cases successful? Such questions will be taken up in follow-on work with the implementer community.

References

Archer, Phil. (2013) Study on persistent URIs: with identification of best practices and recommendations on the topic for the Member States and the European Commission. Retrieved May 02, 2014, from http://philarcher.org/diary/2013/uripersistence .

Bermès, Emmanuelle. (2006). Des identifiants pérennes pour les ressources numériques. Retrieved May 02, 2014, from http://2007.jres.org/planning/pdf/163.pdf.

Berners-Lee, Tim. (1998). Cool URIs don't change. Retrieved May 02, 2014, from http://www.w3.org/Provider/Style/URI.

BnF. (2013). URI and URL in data.bnf.fr. Retrieved May 02, 2014, from http://data.bnf.fr/en/semanticweb#Ancre3.

Byrnes, Margaret. (2000). Defining NLM's Commitment to the Permanence of Electronic Information. ARL 212:8-9. Retrieved May 07, 2014, from http://www.arl.org/newsltr/212/nlm.html

PREMIS Maintenance Activity. (2012). SPAR – Scalable Preservation and Archiving Repository; Retrieved May 02, 2014, from http://www.loc.gov/standards/premis/registry/premis-project_name.php?proj_ID=697.

CDL. (2013). ARK (Archival Resource Key) Identifiers. Retrieved May 02, 2014, from https://wiki.ucop.edu/display/Curation/ARK.

CDL. (2014). Registered Name Assigning Authority Numbers. Retrieved August 14, 2014, from http://www.cdlib.org/uc3/naan_table.html.

Hilse, Hans Werner, and Jochen Kothe. (2006). Implementing Persistent Identifiers. Consortium of European Research Libraries and European Commission on Preservation and Access. Retrieved May 02, 2014, from http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

IETF. (2013). The ARK Identifier Scheme. Internet-Draft. Retrieved May 02, 2014, from http://datatracker.ietf.org/doc/draft-kunze-ark.

Kunze, John. (2003). Towards Electronic Persistence Using ARK Identifiers. California Digital Library. Retrieved May 02, 2014, from https://wiki.ucop.edu/download/attachments/16744455/arkcdl.pdf.

Kunze, John and Adrian Turner. (2010). The ARK Identifier Scheme. Retrieved August 14, 2014, from http://dublincore.org/groups/kernel/spec/.

W3C. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved May 02, 2014, from http://www.ietf.org/rfc/rfc3986.txt.

W3C. (2008). Cool URIs for the Semantic Web. Retrieved May 02, 2014, from http://www.w3.org/TR/cooluris/.

Requirements on RDF Constraint Formulation and Validation

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

Kai Eckert
Research Group Data and Web
Science
University of Mannheim, Germany
kai@informatik.uni-mannheim.de

Abstract

For many RDF applications, the formulation of constraints and the automatic validation of data according to these constraints is a much sought-after feature. In 2013, the W3C invited experts from industry, government and academia to the RDF Validation Workshop, where first use cases have been presented and discussed. In collaboration with the W3C, a working group on RDF Application Profiles (RDF-AP) is currently established in the Dublin Core Metadata Initiative that follows up on this workshop and addresses among others RDF constraint formulation and validation.

In this paper, we present a database of requirements obtained from various sources, including the use cases presented at the workshop as well as in the RDF-AP WG. The database, which is openly available and extendible, is used to evaluate and compare several existing approaches for constraint formulation and validation. We present a classification and analysis of the requirements, show that none of the approaches satisfy all requirements and aim at laying the ground for future work, as well as fostering discussions how to close existing gaps.

Keywords: RDF validation; RDF constraint formulation; RDF constraint validation; requirements; OWL 2; RDF; linked data; semantic web.

1. Introduction

The notion of Linked (Open) Data and its principles clearly increased the acceptance – not to say the excitement – of data providers for the underlying Semantic Web technology. Early concerns of the data providers regarding stability and trustability of the data have been addressed and largely been solved, not only by technical means regarding versioning and provenance, but also by the providers getting accustomed to the open data world with its pecularities.

Linked Data and RDF, however, still are not the primary means to create, store, and manage data on the side of the providers. Linked Data is mostly provided as a view on data, a one-way road, disconnected from the internal data representation. To the obstacles for full adoption of RDF, possibly comparable to XML, belong the lack of accepted ways to formulate (local) constraints on data and to validate data. The W3C reports a consensus among 27 participants from industry, government and academia of RDF Validation Workshop[1] that there are the following needs:

1. Declarative definition of the structure of a graph for validation and description.
2. Extensible to address specialized use cases.
3. A mechanism to associate descriptions with data.

Several use-cases with requirements have been presented at the workshop, further requirements are described in talks about general approaches and experiences outside of RDF, like Dublin Core Application Profiles or XML Schema Definitions. An important finding is that there are non-functional requirements for data validation in a Linked Data setting, particularly the need to

[1] RDF Validation Workshop – Practical Assurances for Quality RDF Data. 10-11 September 2013, Cambridge, MA, USA. http://www.w3.org/2012/12/rdf-val/report

"communicate the constraints against which data is to be validated in a way which is both easy to understand by human beings and discoverable by programs."

SPARQL and SPIN are powerful and widely used for constraint formulation and validation (Fürber and Hepp, 2010), but constraints formulated as SPARQL queries are not as understandable as one wishes them to be. Consider the following example of the simple constraint stating that only dogs are allowed as pets:

```
SELECT ?this ?subope ?object WHERE {
     ?C owl:allValuesFrom :Dog .
     ?C owl:onProperty :hasPet .
     ?C a owl:Restriction .
     ?this rdf:type ?subC . ?subC rdfs:subClassOf* ?C .
     ?this ?subOPE ?object . ?subOPE rdfs:subPropertyOf* :hasPet .
     FILTER NOT EXISTS { ?object rdf:type :Dog . } }
```

This query checks the constraint and returns violating triples, but the actual constraint could be formulated much shorter, for instance using the OWL 2 Functional-Style syntax:

```
SubClassOf( :strictDogOwner ObjectAllValuesFrom( :hasPet :Dog ) )
```

Similarly, but even shorter, as Shape Expression:

```
<StrictDogOwnerShape> { :hasPet :Dog+ }
```

Partly as follow-up to the W3C workshop and partly due to further expressed requirements at the Semantic Web in Libraries conference 2013[2], the Dublin Core Metadata Initiative in collaboration with the W3C currently establishes a Working Group for RDF Application Profiles (RDF-AP WG) that will investigate existing approaches and best-practices, identify possible gaps and propose practical solutions for the representation of application profiles, including the formulation of data constraints[3]. The RDF-AP WG bases its work on currently 8 case studies and use cases provided by internal and external stakeholders, mostly from the library domain. In a heterogeneous environment like the Web, there is not necessarily a one-size-fits-all solution, especially as existing solutions should rather be integrated than replaced, not least to avoid long and fruitless discussions about the "best" approach.

Our work presented in this paper is supposed to lay the ground for subsequent activities in the working group. Our contributions are two-fold: first, we propose to relate existing solutions to specific case-studies and use-cases by means of requirements extracted from the latter and fulfilled by the former. We therefore created and present an exhaustive database of all requirements identified in the validation workshop and the RDF-AP WG. Additionally, we added requirements from other sources, particularly in the form of constraint types that are supported by existing approaches, e.g., expressible in OWL2.

Second, we use this database to provide an overview on different classes of requirements and give examples, to what degree these classes of requirements are supported by different approaches. We want to highlight strengths and weaknesses of these approaches and identify gaps and possible solutions for their elimination.

[2] SWIB13 – Semantic Web in Libraries, 25 - 27 November 2013, Hamburg, Germany. http://swib.org/swib13/

[3] http://wiki.dublincore.org/index.php/RDF-Application-Profiles

2. From a Case Study to a Solution (and Back)

In the development of standards, as in software, case studies and/or use cases are usually taken as starting point. In case studies, the full background of a specific scenario is described, where the standard or the software is to be applied. Use cases are smaller units where a certain action or a typical user enquiry is described. They can be extracted from and thus linked to case studies, but often they are defined directly.

Requirements are extracted from use cases; they form the basis for development and are used to test the result. We specifically use the requirements to evaluate existing approaches for constraint formulation and validation. Via the requirements, the approaches get linked to use cases and case studies and it becomes visible which approaches can be used in a given scenario and what drawbacks might be faced.

We classify the requirements to provide a high-level view on different approaches and to facilitate a better understanding of the problem domain. Our database is openly available and can be extended with new case studies, use cases, requirements and approaches.

Table 1 shows an excerpt from our database. The general structure is a polyhierarchy from case-studies over use-cases and requirements to solutions. All instances contain at least uplinks to the next level, i.e., solutions are linked to requirements that they fulfill and possibly requirements that they explicitly do not fulfill. Requirements are linked to use-cases, which are linked to case studies.

TABLE 1: Database Examples

ID	Title	Links	Description
Case Studies			
CS-1	DPLA	UC-1	The Digital Public Library of America maintains an access portal to digitized cultural heritage objects... We harvest data using several different methods...[4]
Use Cases			
UC-1	Recommended Property	CS-1	Some properties may not be mandatory, but may be recommended to indicate a "value-added" level of compliance with MAPv3...
Requirements			
R-1	Optional Properties	UC-1	A property can be marked as optional. Valid data MAY contain the property.
R-2	Recommended Properties	UC-1, R-3	An optional property can be marked as recommended. A report of missing recommended properties is generated. Fulfilled if R-3 is fulfilled.
R-3	Classified Properties	UC-1	A custom class like "recommended" or "deprecated" can be assigned to properties and used for reporting.
Solutions			
S-1	ShEx	R-1/2/3	Fulfilled: R-1 (minimum cardinality = 0, maximum cardinality = 1). Not fulfilled: R-2, R-3.
S-2	SPIN	R-1/2/3	Fulfilled: R-1, R-2, R-3.

The polyhierarchy allows the linking of all elements to more than one parent, requirements particularly are linked to several use cases. Our goal is to maintain a set of distinct requirements. Only this way it is possible to evaluate the solutions regarding their suitability for the use cases and case studies in our database. Use cases can be shared between case studies as well, but this is harder to maintain as use cases are less formal and often more case specific than a requirement.

[4] http://wiki.dublincore.org/index.php/DPLA_RDF_application_profile_use_cases

Requirement R-2 is an example, where a link between requirements is established. In this case, the link is used to point to a requirement that is "broader" than this requirement, i.e., should that requirement be fulfilled, then this requirement is automatically fulfilled as well. In a similar way requirements can be linked to duplicates if they should occur. Our goal is a relative stability regarding the requirements, which then can prove useful to mediate between data and solution providers.

The database is made available at `http://purl.org/net/rdf-validation`. The initial database was created manually and forms the basis of this paper. The web application to access the database is currently in a beta state and still under development. Nevertheless, the full database can already be browsed online and interested participants can register and contribute to the database.

3. Related Work

Requirements engineering is recognized as a crucial part of project and software development processes. Similar to our collaborative effort, Lohmann et al. propose social requirements engineering, i.e. the use of social software like wikis to support collaborative requirements engineering (Lohmann et al., 2009). Their approach focuses on simplicity and supports in particular the early phases of requirements engineering with many distributed participants and mainly informal collaboration. They emphasize the social experience of developing requirements for software systems: Stakeholders are enabled to collaboratively collect, discuss, improve, and structure requirements. Under the supervision of experts, the requirements are formulated in natural language and are improved by all participants step by step. Later on, experienced engineers may clean and refine requirements. As basis for their work, they developed a generic approach (Softwiki) using semantic technologies and the SWORE ontology for capturing requirements relevant information semantically (Lohmann et al., 2008). The SWORE ontology, as well as a prototypical implementation of their approach is available online[5]. We evaluated the implementation and the ontology regarding a possible reuse, but it turned out that Softwiki focuses clearly on the requirements within a traditional software development process, while we need a broader view including case studies, use cases and various implementing approaches. Nevertheless we will reuse parts of the SWORE ontology and include links wherever possible.

To the best of our knowledge, there is no comparable prior work regarding the collection of a comprehensive list of requirements for the formulation and validation of constraints, neither exist general approaches to compare different solutions based on common or differing requirements. More related work focuses on specific constraint languages and implementations, which we will introduce in the next section.

4. Approaches for Constraint Formulation and Validation

In this section, we present current approaches for constraint formulation and validation which have been the most discussed in the mentioned workshops and WGs. These approaches differ in 2 dimensions: (1) the used constraint language and (2) if they offer validation systems.

OWL, Resource Shapes (ReSh), Shape Expressions (ShEx), Description Set Profiles (DSPs), SPARQL, and SPIN are the most promising and applied constraint languages. Stardog ICV, Pellet ICV, and SPIN use OWL 2 constructs to formulate constraints. SPIN[6] provides a vocabulary to represent SPARQL queries as RDF triples and uses SPARQL to specify inference rules and logical constraints (Fürber and Hepp, 2010). The Pellet Integrity Constraint Validator (ICV)[7] is a proof-of-concept extension for the OWL reasoner Pellet. Stardog ICV[8] validates RDF

[5] http://softwiki.de/netzwerk/en/

[6] http://spinrdf.org

[7] http://clarkparsia.com/pellet/icv/

[8] http://docs.stardog.com/icv/icv-specification.html

data stored in a Stardog RDF database. ReSh[9] defines its own RDF vocabulary Open Services for Lifecycle Collaboration (OSLC) to define constraints (Ryman et al., 2013). ShEx[10] also specifies a new constraint language whose syntax and semantics are similar to regular expressions. DCMI RDF Application Profile (AP)[11] and Bibframe[12] are approaches to specify profiles for application-specific purposes. DCMI RDF-AP uses DSP[13] as generic constraint language which is also intuitive for non-experts. The Bibframe constraint language has a strong overlap with DSP. Kontokostas et al. define 17 data quality integrity constraints represented as SPARQL query templates called Data Quality Test Patterns (DQTP) (Kontokostas et al., 2014). Schemarama[14] is based on the Squish RDF language instead of SPARQL. For XML, Schematron[15] is an ISO standard for validation and quality control of XML documents based on XPath and XSLT. XML Schema[16] is the primary technology for specifying and constraining the structure of XML documents.

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

In this paper, we evaluate to which extend these approaches cover classes of requirements (1) to express different types of constraints and (2) to formulate constraints. For the formulation of constraints, it is important that the constraint language is concise and intuitive and that the declarative constraint language is translated to an implementation language like SPARQL in order to execute constraint validation automatically. In form of concrete examples, we show how current approaches can be used to express different types of constraints and how they can be used together to fulfill the majority of the identified requirements classes.

5. Requirements

Use cases discussed within the scope of the mentioned workshops and working groups led to the definition of requirements on RDF constraint formulation and validation. We classified these requirements into the 2 top-level categories ’Constraint Formulation’ and ’Constraint Expressivity’.

5.1. Formulation of Constraints

Intuitive and concise language. We claim that all constraints can be expressed using the low-level language SPARQL. The majority of the constraints can also be written more declaratively, intuitively, and concisely in form of OWL 2 axioms in the concrete syntax Turtle. Although, OWL 2 is a very expressive language, we cannot express every constraint in OWL 2. The succeeding existential quantification contains those individuals that are connected by the `:fatherOf` property to individuals that are instances of the class `:Man`. The ontology, the constraint, and RDF data are expressed with the same OWL 2 axiom and the same concrete syntax:

```
[ rdfs:subClassOf [
   a owl:Restriction;
   owl:onProperty :fatherOf;
   owl:someValuesFrom :Man ] ] .
```

[9] http://www.w3.org/Submission/shapes/

[10] http://www.w3.org/2013/ShEx/Definition

[11] http://dublincore.org/documents/singapore-framework/

[12] http://bibframe.org/

[13] http://dublincore.org/documents/dc-dsp/

[14] http://swordfish.rdfweb.org/discovery/2001/01/schemarama/

[15] http://www.schematron.com/

[16] http://www.w3.org/TR/xmlschema-1/

The main purpose of OWL 2 is to infer new knowledge from existing schemata and data rather than to check data for inconsistencies. Therefore, most constraint validation approaches define constraints with other high-level declarative languages, even though most people are familiar with OWL 2 and its concise human-understandable concrete syntax Turtle. OWL 2 can be used to describe RDF data, to infer new knowledge, and to validate RDF data using the same expressive OWL 2 axioms. With XML Schemas, we also structure and validate our data according to that structure.

Shape Expressions contain elements from regular expressions making the language concise and intuitive. In the following example, an employee has at least 1 given name, 1 family name, any number of phone numbers, and 1 mail box:

```
<EmployeeShape> {
     foaf:givenName xsd:string+ ,
     foaf:familyName xsd:string ,
     foaf:phone IRI* ,
     foaf:mbox IRI }
```

As different constraints can be expressed with different languages, we propose to use multiple languages to define constraints depending on the requirements which have to be satisfied.

Translated to implementation language. High-level declarative languages like OWL 2 cannot be executed directly to validate constraints. Therefore, we take a low-level execution language like SPARQL. Sirin and Tao (2009) showed how constraints can be translated to nonrecursive Datalog programs for validation and Angles and Gutierrez (2008) explained that SPARQL has the same expressive power as nonrecursive Datalog programs. As a consequence, we can also use SPARQL queries to validate constraints. Thus, constraint validation can be reduced to SPARQL query answering. The participants of the 2013 W3C RDF Validation workshop agreed that SPARQL should be the language to execute constraint validation[17]. Furthermore, all evaluated constraint validation approaches execute constraint validation with SPARQL. The next SPARQL query shows how the OWL 2 existential quantification is implemented in SPIN:

```
CONSTRUCT {
     _:violation
         a spin:ConstraintViolation ;
         rdfs:label ?violationMessage
         spin:violationRoot ?this }
WHERE {
     ?this rdf:type ?subC . ?subC rdfs:subClassOf* ?C .
     ?C owl:someValuesFrom ?CE .
     ?C owl:onProperty ?OPE .
     ?C a owl:Restriction .
     FILTER ( sp:not ( spl:hasValueOfType ( ?this, ?OPE, ?CE ) ) ).
     FILTER EXISTS { ?this ?OPE ?object . ?object rdf:type owl:Thing . }
     BIND ( ( ... ) AS ?violationMessage ) . }
```

RDF representation of constraints. One of the main benefits of SPIN is that arbitrary SPARQL queries and thus constraints are represented as RDF triples. SPIN provides a vocabulary, the SPIN SPARQL Syntax, to represent SPARQL queries in RDF. The benefits of an RDF representation of constraints are:

- constraints can be consistently stored together with ontologies and RDF data

[17] http://www.w3.org/2013/09/10-rdfval-minutes

- constraints can be easily shared on the web of data
- constraint validation can be executed automatically
- constraints can be processed by a plethora of already existing RDF tools
- constraints are linked to RDF data

The subsequent code snippet demonstrates how SPIN represents SPARQL 1.1 NOT EXISTS filter expressions in RDF:

```
FILTER NOT EXISTS { ?person foaf:name ?name }
-----
[   a sp:Filter ;
    sp:expression [
        a sp:notExists ;
        sp:elements (
            [   sp:subject [ sp:varName "person" ] ;
                sp:predicate foaf:name ;
                sp:object [ sp:varName "name" ] ] ) ] ] )
```

Our approach, which is implemented in Java, executes constraint validation with SPIN. SPIN templates define the validation of both OWL 2 constraints and constraints only expressible with SPARQL. These constraints are checked for each resource of the type `owl:Thing` (all resources are assigned to the common super-class `owl:Thing`).

Constraint validation results. Like ontologies, instance data, and constraints, we should also represent constraint violations in RDF. SPIN templates construct (SPARQL CONSTRUCT) constraint violation triples containing information about constraint violations, which cannot be expressed directly in OWL 2:

```
CONSTRUCT {
     _:icViolation
         a spin:ConstraintViolation ;
         rdfs:label ?violationMessage ;
         spin:violationRoot ?violationRoot ;
         spin:violationPath ?violationPath ;
         spin:violationSource ?violationSource ;
         spin:fix ?violationFix ;
        :severityLevel ?severityLevel }
```

Constraint violations (of the type `spin:constraintViolation`) should provide a useful message (`rdfs:label`) explaining the reasons why the data did not satisfy the constraints, which aids data debugging and repair. If we do not state the triples `:Peter :fatherOf :Stewie .` and `:Stewie a :Man .`, the SPIN template checking the OWL 2 existential quantification on the object property `:fatherOf` constructs a constraint violation triple raising the message 'ObjectSomeValuesFrom(:fatherOf :Man) - :Stewie must be an instance of :Man'. Now, you know exactly why the data violated this constraint and you know where you have to modify your data. Constraint violation triples contain references to triples causing the constraint violations (`spin:violationRoot`) and references to constraints causing constraint violations (`spin:violationSource`). In our example, the subject `:Peter` causes the constraint violation and the constraint `:ObjectSomeValuesFrom` constructs the constraint violation triple. To fix constraint violations we need to give some guidance how to become valid data (`spin:fix`). Appropriate triples may point to useful messages explaining in detail how to overcome constraint violations. Constraint violations can be classified according to different levels of severity (`:severityLevel` having controlled vocabulary as range with elements like `:Error` and

:Warning). It is also important to find not validated triples, i.e. triples which have not been validated by any constraint, as it may be enforced that every triple of the data have to be validated.

5.2. Constraint Expressivity

Cardinality Restrictions. Class expressions in OWL 2 can be formed by placing restrictions on the cardinality of object and data property expressions. All cardinality restrictions can be qualified or unqualified. The class expressions contain those individuals that are connected by a property expression to at least, at most, and exactly a given number of instances of a specified class expression. Qualified and unqualified cardinality restrictions can be expressed in OWL 2:

```
:CE rdfs:subClassOf [
     a owl:Restriction ;
     owl:maxQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
     owl:onProperty :hasSon ;
     owl:onClass :Man ] .
:Peter a :CE ;
     :hasSon :Stewie [ a :Man ] .
```

:Peter is an instance of the class expressions containing those individuals having at most 1 son which is :Stewie in the RDF instance data. If we state that :Peter has a second son or if we do not assign :Stewie to the class :Man, the qualified maximum cardinality restriction will be violated. SPIN, Stardog, and Shape Expressions are the only approaches with which qualified and unqualified cardinality restrictions on data and object properties can be specified.

Disjointness. Disjointness of classes and union of class expressions, (class-specific) object and data properties, and individuals is a very important type of constraints which can be completely covered with SPIN (implementing OWL 2 constructs). An OWL 2 disjoint union axiom DisjointUnion(C CE_1 ... CE_n) states that a class C is a disjoint union of the class expressions CE_i, $1 \leq i \leq n$, all of which are pairwise disjoint. Each instance of C is an instance of exactly one CE_i, and each instance of CE_i is an instance of C[18]. According to the next disjoint union of 2 class expressions, each child is either a boy or a girl, each boy is a child, each girl is a child, and nothing can be both a boy and a girl. As in this example, :Stewie is both a boy and a girl, a constraint violation is raised:

```
:Child owl:disjointUnionOf ( :Boy :Girl ) .
:Stewie a :Child ; a :Boy ; a :Girl .
```

Disjoint groups of object and data properties can be expressed in OWL 2:

```
[   rdf:type owl:Class ;
    owl:unionOf (
        [   rdf:type owl:Restriction ;
             owl:qualifiedCardinality 1  ;
            owl:onProperty foaf:name ;
            owl:onClass xsd:string ]
        [   rdf:type owl:Class ;
            owl:intersectionOf (
                [   rdf:type owl:Restriction ;
                    owl:minQualifiedCardinality 1 ;
                    owl:onProperty foaf:givenName ;
                    owl:onClass xsd:string ] .
```

[18] http://www.w3.org/TR/owl2-syntax/

```
                [   rdf:type owl:Restriction ;
                    owl:qualifiedCardinality 1 ;
                    owl:onProperty foaf:familyName ;
                    owl:onClass xsd:string ] ) ] ) ] .
```

In this example, we define a shape for persons. A person has either a FOAF name or 1 or more given names and 1 family name. Although this kind of constraint can be realized in OWL 2, the definition of disjoint groups of properties is not that intuitive and declarative. Exactly the same constraint can be expressed more concisely with Shape Expressions:

```
<PersonShape> {
   ( foaf:name xsd:string
     |
     foaf:givenName xsd:string+ ,
     foaf:familyName xsd:string ) }
```

Shape Expressions and SPIN are the only approaches to specify disjoint groups of properties for given classes.

Constraints on RDF Properties. Object as well as data properties may be constrained. The main component of an OWL 2 ontology is a set of axioms - statements that say what is true in the domain. OWL 2 provides axioms that can be used to characterize and establish relationships between object and data property expressions. An object property functionality axiom states that an object property expression is functional - that is, for each individual x, there can be at most one distinct individual y such that x is connected by the object property expression to y[19]. With Pellet ICV, we can state a couple of object and data property axioms like the following object property functionality axiom in OWL Turtle syntax (Sirin and Tao, 2009):

```
:isManufacturedBy a owl:FunctionalProperty .
:Product :isManufacturedBy :Manufacturer1 , :Manufacturer2 .
```

The object property `:isManufacturedBy` is defined as functional. The OWL interpretation would infer that the manufacturers are the same resources, as nothing contradicts the inference that these two manufacturers are the same and there is no Unique Name Assumption. With constraint semantics, however, a constraint violation is raised. With Resource Shapes 2.0 and Shape Expressions it is not possible to declare functionality axioms on object and data properties. We can define these axioms with SPIN (and OWL 2), Stardog, and Pellet.

Object property paths (supported by Stardog and SPIN) are important constraints within various domains. Object property chains can be expressed as OWL 2 axioms SubObjectPropertyOf(ObjectPropertyChain(OPE$_1$ ... OPE$_n$) OPE) stating that, if an individual x is connected by a sequence of object property expressions OPE$_1$, ..., OPE$_n$ with an individual y, then x is also connected with y by the object property expression OPE[20]. As the triple `:Stewie :hasAunt :Carol .` is not contained in the following data set, a constraint violation results:

```
:hasAunt owl:propertyChainAxiom ( :hasMother :hasSister ) .
:Stewie :hasMother :Lois . :Lois :hasSister :Carol .
```

[19] http://www.w3.org/TR/owl2-syntax

[20] http://www.w3.org/TR/owl2-syntax

Constraints on RDF objects. For RDF objects, we can state constraints such as allowed values, default values, and negative object constraints. Resource Shapes 2.0 enables defining allowed values for RDF objects as well as RDF literals:

```
:oslc-change-request a oslc:ResourceShape ;
   oslc:property :oslc_cm-status .
:oslc_cm-status a oslc:Property ;
   oslc:allowedValues :status-allowed-values .
:status-allowed-values a oslc:AllowedValues ;
   oslc:allowedValue "Done" , "InProgress" , "Submitted" .
```

The constraint above specifies the only allowed values of the status data property for change request resources. If change requests have other status values, constraint violations will be raised. In addition to Resource Shapes 2.0, the DCMI RDF-APs and SPIN (and OWL 2) allow specifying allowed values for RDF literals. For RDF objects, we can apply the approaches Resource Shapes 2.0, Shape Expressions, DCMI RDF-APs, and SPIN (and OWL 2) to define allowed values.

With DCMI RDF-APs and SPIN, we can declare that RDF objects and literals have to be part of specific controlled vocabularies. These statements are represented with DCMI RDF-APs using an RDF triple comprising an RDF subject that is the value RDF node, an RDF predicate `dcam:memberOf`, and an RDF object with a corresponding RDF URI Reference being the DCAM vocabulary encoding scheme URI[21]. The following excerpt states that a given book is assigned to the topic 'Ornitology' which is part of a particular controlled vocabulary:

```
:Book
     dcterms:subject [
         rdf:value "Ornitology" ;
         dcam:memberOf :ControlledVocabulary ] .
```

Constraints on RDF Literals. Constraint on RDF literals are not that significant in the Linked Data community, but they are very important in communities like the library domain. For RDF literals, range-specific, constraining facet-specific, datatype-specific constraints, and language-specific can be defined. We can restrict the datatypes, RDF literals have to correspond to, with XML Schema constraining facets. SPIN allows us to implement all constraining facets. DQTPs enables constraining literal values to match or not to match a certain regex pattern (`xsd:pattern`):

```
SELECT DISTINCT ?s WHERE { ?s %% P1 %% ? value .
     FILTER ( %% NOP %% regex (str (? value ), %% REGEX %) ) }
```

P1 is the property we need to check against REGEX and NOP can be a not operator (!) or empty. An example binding could be to check if the dbo:isbn format is different (!) from “ˆ([iIsSbBnN 0-9-])*$” (Kontokostas et al., 2014). DQTPs also enables constraining literal values (having a certain datatype) to be or not to be within a specific range (`xsd:maxInclusive, xsd:maxExclusive, xsd:minExclusive, xsd:minInclusive`):

```
SELECT DISTINCT ?s WHERE {
     ?s rdf:type %% T1 %% . ?s %% P1 %% ?value .
     FILTER ( %% NOP %% (?value < %% Vmin %% || ?value > %% Vmax %%))) }
```

[21] http://dublincore.org/documents/dc-rdf/

For instance, we can restrict geographical longitudes and latitudes (`geo:lat, geo:long`) of a spatial feature to be within the range [-90,90] (Kontokostas et al., 2014). Furthermore, we implemented the constraining facet `xsd:whiteSpace` in SPIN to avoid leading and trailing white spaces in literals. Sub-types of language-specific constraints on RDF literals are constraints (1) to check if a literal for a specific data property within the context of a particular class has a given language tag, (2) to check whether the literal, within the context of a given property and class, is missing, or (3) to ensure that resources of a given type must have at most 1 value of a specific language for a given data property (e.g. a single English ("en") `rdfs:label`). Default values can be defined with Bibframe, Resource Shapes 2.0, and SPIN. For this purpose, SPIN constructors may contain SPARQL CONSTRUCT queries for specific classes (e.g. USA is the birth country of each `USCitizen`):

```
:USCitizen a rdfs:Class ;
     spin:constructor [ a sp:Construct ; sp:text """
     CONSTRUCT { ?this :birthCountry "USA" . } WHERE {} """ ] .
```

6. Evaluation

In this section, we evaluate current approaches according to the top-level classification of constraint validation requirements. This kind of evaluation is crucial for future improvements regarding constraint formulation and validation of both existing and new approaches. The underlying facts result primarily from the individual official specifications. We categorize requirements classes to see which requirements are well, badly, and limited satisfied by which approaches. The goal of this evaluation is not to completely evaluate all currently available constraint validation approaches. We want to show in a generic way that none of the current approaches satisfies all requirements and that different approaches cover different requirements classes. Case studies and use cases define what requirements classes have to be covered. This evaluation indicates which approaches to use to cover specific requirements classes and therefore use cases. There are 2 first level requirements classes: 'constraint expressivity' and 'constraint formulation'. Tables 2 and 3 show for each approach what second level requirements classes are covered to which extent. Numbers in brackets behind requirements classes indicate the number of requirements contained in that class. Numbers in brackets in table cells indicate that requirements are limited satisfied.

TABLE 2: Constraint Expressivity

Requirements Classes	BF	DCMI	DQTP	Pellet	RS	SE	SPIN	Stardog
Disjointness (8)	✗	✗	3		✗	3	5	
Equivalence (4)	✗	✗			✗	✗	4	
Constraints on RDF properties (20)			12	3	1(1)	2	20	7
Constraints on RDF objects (7)	2	2	1	1	3(1)	5	5	2
Constraints on RDF literals (14)	2	2	4		3(1)	2(1)	7	
Identification (5)	✗	✗			✗	(2)	4	✗
Uniqueness (2)							1	
Provenance Constraints	✗	✗	✗	✗	✗	✗	✗	✗
Constraints on Individuals (6)	✗	✗			✗	✗	6	
Class Relationships (4)			2			1	4	1
Set-Oriented Operations (6)				2			6	3
Property Occurrences (9)	1	1	1		3	6	6	2
Property Restrictions (10)			1	2		2	8	3
Cardinality Restrictions (12)	✗	✗	6	✗	(12)	12	12	3

Good Coverage. Although equivalence (e.g. equivalent classes) is only considered by 1 approach (SPIN), all 4 associated requirements are satisfied. 1 approach (SPIN) covers all 20 requirements on RDF properties constraints (e.g. object property paths) and 2 approaches (DQTP and Stardog) fulfill half of these requirements. Just 1 approach (SPIN) covers 4 of 5 identification requirements (e.g. to check if IRIs correspond to specific patterns). Class expressions represent sets of individuals by formally specifying conditions on the individuals' properties; individuals satisfying these conditions are said to be instances of the respective class expressions. Subcategories of this requirements class are well satisfied by 3 approaches (DQTP, Shape Expressions, and Stardog) and nearly exhaustively satisfied by 1 approach (SPIN). Class-relationships (e.g. subsumption) and set-oriented operations (e.g. negation of classes) are not supported by many approaches. In contrast, property occurrences (e.g. mandatory or optional), property restrictions (e.g. existential quantifications), and cardinality restrictions are supported by the majority of current approaches. Constraints on individuals (e.g. negative object property assertions) are only considered by 1 approach (SPIN) which fulfills all associated requirements.

Limited Coverage. Approach developers should mention requirements which are not covered exhaustively by current approaches. Only 3 approaches (DQTP, Shape Expressions, and SPIN) consider disjointness constraints (e.g. class-specific disjoint property groups) and 1 approach (SPIN) covers 5 of 8 disjointness requirements. 5 of 7 requirements on RDF objects constraints (e.g. allowed values) can be expressed with 2 approaches (Shape Expressions and SPIN). There are 2 requirements to ensure uniqueness (e.g. unique URIs), but only 1 approach (SPIN) satisfies 1 requirement. Other approaches do not cover uniqueness requirements.

Bad Coverage. For future development of approaches it is crucial to especially consider requirements which are currently not satisfied at all by any approach. So far, provenance constraints are not considered by approach developers. Most approaches satisfy just 2 of 14 requirements on RDF literal constraints (e.g. range of literal values). At least 1 approach (SPIN) covers 50% of these requirements.

Table 3 shows constraint formulation requirements (classes) and their coverage by current approaches. Even though, almost each constraint language is intuitive, only 4 constraint languages can be seen as both intuitive and concise (Pellet, Shape Expressions, SPIN, and Stardog). 3 of these 4 approaches use OWL 2 as declarative language - the standard language to define ontologies. Shape Expressions uses a language similar to regular expressions.

TABLE 3: Constraint Formulation

Requirements Classes	BF	DCMI	DQTP	Pellet	RS	SE	SPIN	Stardog
Intuitive Language	✓	✓	~	✓	✓	✓	✓	✓
Concise Language	✗	✗	✓	✓	✗	✓	✓	✓
Translated to Implementation Language	✗	✗	✓	✓	✗	✓	✓	✓
Implemented Constraint Validation	✗	✗	✓	✓	✗	✓	✓	✓
Implementation Publicly Available	✗	✗	✓	~	✗	✗	✓	~
RDF Representation of Constraints	✗	✗	✗	✗	✗	✗	✓	✗
Constraint Validation Results (10)	✗	✗	2	2	✗	6	9	1

Five of 8 approaches translate declarative constraints formulations to an implementation language (e.g. SPARQL) to execute constraint validation. It is very important for future enhancements by the whole community that implementations are not only existent but also publicly available. 5 of 8 approaches are implemented, but implementations are publicly available for only 2 approaches (public availability of implementations is limited for 2 further approaches). Constraints are represented as RDF triples by only 1 approach (SPIN). RDF should be the natural and standard format to represent constraints within the Linked Data community. 2 approaches (Shape Expressions and SPIN) cover almost all requirements on validation results (e.g. providing some guidance how to become valid data). Unfortunately, 3 of the remaining approaches cover requirements on validation results very poorly.

7. Conclusion and Future Work

Heterogeneous approaches with different strengths and weaknesses are not a bad thing; we do not expect there to be a one-size-fits-all solution, nor do we aim at creating one. With this paper, we rather want to raise the awareness towards the differences and commonalities of existing approaches as well as to shed some light on the different requirements that data providers currently have. Therefore, we presented our approach to collect case studies, use cases and especially requirements collaboratively and in structured form. By linking the requirements to existing constraint languages and validation systems, we could identify strengths and weaknesses, commonalities and differences not only intellectually, but based on reliable data.

The main purpose of this work is to support discussions of the different approaches and to help stakeholders in the choice or in the development of appropriate solutions. In the context of application profiles, where the publication of constraints together with the data model is crucial, we want to emphasize the need for concise, easy to understand constraint languages. This requirement is often neglegtected in discussions of approaches. While consistency is understandably desired, it has to be questioned if one constraint language can fulfill all requirements without being overly complicated or if different approaches should rather be used for different classes of requirements. This holds especially for different levels of abstraction, as the possibility to define constraints on the format of RDF literals compared to constraints on the availability or special properties of provenance information. Both represent examples where all current approaches lack proper support.

Gaps within a class of requirements, e.g., disjointness, constraints on RDF objects, or uniqueness, should be easier to close within the existing approaches. This would lead to a harmonization of the approaches regarding their expressivity and enable translations in-between or towards a general constraint language, e.g., the translation of well-readable constraints in any language to executable SPARQL queries. The latter is especially promising considering that SPARQL is able to fulfil all functional requirements and already considered by many as a practical solution to formulate constraints.

As future work, we plan to provide a complete implementation of OWL 2 constraints in form of SPIN templates to demonstrate this approach. We will extend and maintain the requirements database and hope to establish it as an important tool for the advancement of constraint formulation and validation in RDF. Within the DCMI RDF Application Profiles Working Group, we pursue the establishment of application profiles that among others allow to link constraints directly to published datasets and ontologies.

Acknowledgements

Kai Eckert is funded by the European Commission within the DM2E project (http://dm2e.eu)

References

Angles Renzo and Gutierrez Claudio. (2008). The expressive power of SPARQL. In Proceedings of the 7th International Semantic Web Conference (ISWC2008), pages 114–129, 2008.

Fürber Christian and Hepp Martin. (2010). Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In Witold Abramowicz and Robert Tolksdorf, editors, Business Information Systems, volume 47 of Lecture Notes in Business Information Processing, pages 35–46. Springer Berlin Heidelberg, 2010.

Lohmann Steffen, Dietzold Sebastian, Heim Philipp, and Heino Norman. (2009). A web platform for social requirements engineering. In Jürgen Münch and Peter Liggesmeyer, editors, Software Engineering (Workshops), volume 150 of LNI, pages 309–315. GI, 2009.

Lohmann Steffen, Heim Philipp, Auer Sören, Dietzold Sebastian, and Riechert Thomas. (2008). Semantifying requirements engineering – the softwiki approach. In Proceedings of the 4th International Conference on Semantic Technologies (I-SEMANTICS '08), J.UCS, pages 182–185, 2008.

Kontokostas Dimitris, Westphal Patrick, Auer Sören, Hellmann Sebastian, Lehmann Jens, Cornelissen Roland, and Zaveri Amrapali. Test-driven evaluation of linked data quality. In Proceedings of the 23rd International Conference on World

Ryman Arthur G., Le Hors Arnaud, and Speicher Steve. (2013) Oslc resource shape: A language for defining constraints on linked data. In Christian Bizer, Tom Heath, Tim Berners-Lee, Michael Hausenblas, and Sören Auer, editors, LDOW, volume 996 of CEUR Workshop Proceedings. CEUR-WS.org, 2013.

Sirin E. and Tao J.. (2009). Towards integrity constraints. In Proceedings of the Workshop on OWL: Experiences and Directions, OWLED 2009, 2009.

Wide Web, WWW '14, pages 747–758, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

Extracting Description Set Profiles from RDF Datasets using Metadata Instances and SPARQL Queries

Tsunagu Honma Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan tsuna@slis.tsukuba.ac.jp	Kei Tanaka NTT DATA Corporation, Japan telekyon.official@gmail.com	Mitsuharu Nagamori Faculty of Library, Information and Media Science, University of Tsukuba, Japan nagamori@slis.tsukuba.ac.jp	Shigeo Sugimoto Faculty of Library, Information and Media Science, University of Tsukuba, Japan sugimoto@slis.tsukuba.ac.jp

Abstract

A variety of communities create and publish metadata as Linked Open Data (LOD). Users of those datasets find and use them for their own purposes and may combine the datasets to add value. Each LOD dataset uses various vocabularies, structures and constraints for describing resources. In order to improve the usability of LOD datasets, it is very important for metadata designers to enhance the interoperability of their own metadata with that of other datasets. In order to create new interoperable metadata, metadata schema designers have to understand the Application Profiles of the existing LOD datasets.

Dublin Core Description Set Profile (DSP) is a component of Dublin Core Application Profiles. A DSP describes the structures and constraints of metadata in an application (e.g., resource classes, properties cardinality and value scheme). Metadata schema registries, which collect and provide metadata schemas, have a large potential for helping metadata schema designers find, compare, and adopt existing schemas. However, most LOD datasets are not published with their DSPs. As a result, metadata schema designers have to look at each dataset and guess the DSPs.

This paper proposes a method to extract the structural constraints of metadata records automatically from metadata instances using existing metadata schema. The goal of this study is to reduce the cost of metadata schema extraction and to increase the number of metadata schemas registered in metadata schema registries. We have experimentally extracted constraints from LOD datasets using SPARQL. To evaluate, we applied our approach to 10 datasets in the DataHub. By comparing the structural constraints that were extracted using our approach with a manual approach, we found that our approach was able to extract more constraints.

Keywords: application profiles; metadata schema design; metadata schema extraction

1. Introduction

A considerable number of metadata datasets are published as Linked Open Data (LOD)[1] for sharing on the Web. LOD is widespread across many specific domains such as government, geography and e-science. Many communities create and publish LOD datasets on the Web and users are free to combine those datasets. Before designing new LOD datasets, metadata schema designers design a new application profile, which defines some constraints of metadata that are important for users of datasets. Particularly, in order to mash-up different datasets, metadata schema designers should create schema that enhance the interoperability of those metadata.

Application Profiles (Coyle and Baker, 2009) are helpful for users to understand the constraints of datasets. Dublin Core Description Set Profile (DSP) (Nilsson, 2008) is a component of an application profile, which explains the structural constraints of metadata

[1] http://linkeddata.org/

instances (Nilsson and Baker, 2008). If metadata schema designers are able to find and use DSPs, they can understand what vocabularies, structures, and constraints are used for describing datasets in that specific domain.

There are some metadata schema registries for accumulating and publishing metadata vocabularies and application profiles. Metadata schema designers can use those registries for finding existing application profiles that are similar to their own application profile. In order to cover a more specific domain, we have to increase the number of application profiles. However, most LOD datasets are not published with their profiles (Nishide, et al,. 2013). Therefore, metadata schema designers have to look into datasets and try to deduce their structural constraints. There are a lot of datasets in each specific domain, and those datasets are often too large to look into to determine structural constraints. It is therefore costly for metadata schema designers to have to make deductions about structural constraints manually.

We propose a method to extract the structural constraints of LOD datasets automatically. Creators of LOD datasets describe metadata instances based on their implicit or explicit structural constraints. Therefore, we use metadata instances, which are included in LOD datasets and existing metadata schema, for extracting structural constraints. We extract structural constraints from LOD datasets using SPARQL. We create Description Templates for each class membership, which resources are instances of. After creating Description Templates, we also extract property URIs, value types, language tags and datatypes for creating Statement Templates.

We apply our approach in practice to 10 datasets in the DataHub for evaluating our approach and clarifying issues which we need to solve for improving our method.

2. Sharing Application Profiles to Design a New Interoperable Schema

When metadata schema designers design a new application profile, they try to find existing application profiles in order to 1) reduce the cost of designing application profiles, 2) improve the interoperability of their metadata and 3) develop requirements for their metadata. Creating application profiles from scratch comes at a high cost, because metadata schema designers have to find suitable metadata vocabularies and structures for their purposes. If there are existing application profiles which have been created for similar purposes, designers can reuse those schema to reduce the cost of finding metadata vocabularies and deciding on the structure of metadata. As a result, the new application profile has improved interoperability because schema designers reuse common vocabularies and structures in the specific domain in which their metadata is used. Through reusing and customizing existing application profiles, metadata schema designers develop requirements for their metadata

In order to accomplish these goals, metadata schema designers should find and reuse existing application profiles in the same domain. Metadata schema registries are useful for metadata schema designers to find existing parts of application profiles. Metadata schema registries support the sharing of metadata schema on the web and promote reuse of metadata schemas and metadata interoperability (Nagamori et al., 2011). The Open Metadata Registry (Hillmann et al., 2006) is one such metadata schema registry. This registry can store metadata vocabularies and metadata schema in the form of element sets. MetaBridge (Nagamori et al., 2011) is also a metadata schema registry which is compatible with OWL-DSP based on DSP. If metadata creators share their application profile explicitly in those registries, metadata schema designers can use those registries as examples of metadata structures and constraints when they design new application profiles.

The number of application profiles that are registered in those registries is not enough for metadata schema designers to find and reuse those profiles. Therefore, it is important to create and register application profiles of various datasets. If metadata creators publish LOD datasets with their application profiles, schema registries can accumulate and share those application profiles. However, most LOD datasets are published without explicit application profiles. For that reason, one has to look into each LOD dataset and create its application profile manually. LOD datasets are often too large for observing as a whole, and observing those datasets and creating application profiles are difficult for metadata schema designers. It is necessary to extract application profiles from existing LOD datasets automatically.

There is related work in the area of schema extraction (Chidlovskii, 2002). Here, the researchers proposed methods for extracting XML Schema. XML Schema defines the structural constraints of metadata, which have been serialized in XML, such as the hierarchies of each XML element and its attribute. However, we would like to extract the structural constraints of resources, properties and values that are described with the RDF model, not only serialized with XML. Such constraints are independent of the serialization found in XML elements hierarchies. SchemEX (Konrath et al., 2012) is an existing approach for extracting metadata schema from LOD datasets. This approach extract schema that includes RDF type clusters and relationships between resources that are instances of type clusters. Those schema abstract structural constraints about dataset with typed resources and properties, but not define metadata value constraints, especially literal value constraints such as datatypes and language tags.

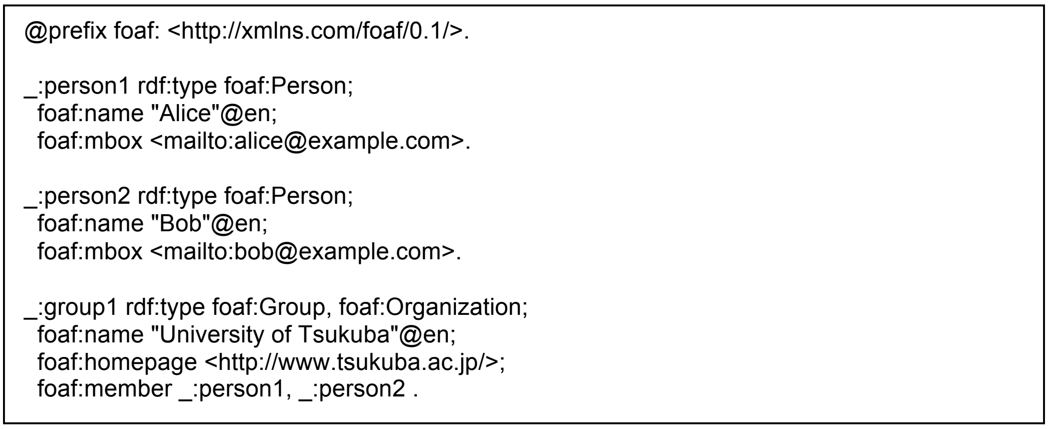
In this research, we propose a method to extract application profiles for LOD datasets automatically using metadata instances and existing schema. In the Singapore Framework, an application profile consists of five components. This research aims to extract Description Set Templates, which define the structural constraints of metadata instances. Metadata instances are described based on implicit or explicit structural constraints. We can extract those constraints from existing metadata instances.

3. Extracting Structural Constraints from Metadata Instances

Definitions of metadata vocabularies, structural constraints of metadata and description formats are all components of a metadata schema. In this research, our goal is to extract structural constraints as a DSP when a user inputs metadata instances. A DSP consists of Description Templates and Statement Templates. Description Templates define the constraints of resources, and Statement Templates define the constraints of attributes. In DSP, we are able to describe the following constraints using Description Templates and Statement Templates.

- Description Templates
  - Resource class membership constraints
  - Statement Templates which belongs to this Description Template
- Statement Templates
  - Property URI
  - Type constraint, "literal" or "non-literal"
  - Class membership of non-literal metadata values
  - Datatypes and language tags of literal metadata values

In this section, we explain our approach for extracting structural constraints with an example. Figure 1 shows an example of metadata instances. The example shows that *_:group1* is an instance of *foaf:Group ∩ foaf:Organization*. This resource has two members using *foaf:member*, *_:person1* and *_:person2* which have their own names and email addresses with *foaf:name* and *foaf:mbox*. Our goal is extracting the structural constraints of these metadata instances as seen in table 1 and table 2. Table 1 shows the constraints of resources which are instances of *foaf:Group ∩ foaf:Organization*. Table 2 shows the constraints of resources which are instances of *foaf:Person*.



```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.

_:person1 rdf:type foaf:Person;
  foaf:name "Alice"@en;
  foaf:mbox <mailto:alice@example.com>.

_:person2 rdf:type foaf:Person;
  foaf:name "Bob"@en;
  foaf:mbox <mailto:bob@example.com>.

_:group1 rdf:type foaf:Group, foaf:Organization;
  foaf:name "University of Tsukuba"@en;
  foaf:homepage <http://www.tsukuba.ac.jp/>;
  foaf:member _:person1, _:person2 .
```

FIG. 1. An example of metadata instances for extracting structural constraints of metadata

TABLE 1: Structural constraints of an instance of foaf:Group ∩ foaf:Organization

Attribute	Property	Value Constraints
name	foaf:name	rdfs:Literal, @en
website	foaf:homepage	foaf:Document
member	foaf:member	foaf:Person

TABLE 2: Structural constraints of an instance of foaf:Person

Attribute	Property	Value Constraints
name	foaf:name	rdfs:Literal, @en
email	foaf:mbox	rdfs:Resource

Metadata instances are described based on the above constraints, and we extract them from metadata instances using the following steps. In each step, we extract resources, properties and values using SPARQL because we need to estimate the structural constraints of metadata instances. Before extracting the structural constraints, we loaded metadata instances in an RDF database.

Step 1: Get the class membership which resources are instances of

Step 2: Get the properties for each class membership

Step 3: Get a value type constraint (literal or non-literal)

Step 4: Get other value constraints

Step 4-1: Get literal value constraints (e.g., language tag and datatype)

Step 4-2: Get non-literal value constraints (e.g., resource class membership and base URI)

In the first step, we extract class memberships of resources which are described using *rdf:type* because typed resources are useful starting anchors for defining Description Templates. In our example, there are two class memberships, *foaf:Person* and (*foaf:Group ∩ foaf:Organization*). We extract those memberships using a SPARQL query, which is shown in Figure 2, and create two Description Templates.

```
SELECT DISTINCT (GROUP_CONCAT(DISTINCT(?type) ; separator = ", ") as ?types)
WHERE {
  ?s rdf:type ?type.
  ?s ?p ?o.
  FILTER(?p!=rdf:type)
}
GROUP BY ?s
ORDER BY ?type
```

FIG. 2. A SPARQL query for extracting the class membership which resources are instances of

The second step is a process for creating Statement Templates. Statement Templates are created for defining the constraints of metadata attributes. In this step, we execute queries to find properties for each class membership, which are defined by Description Templates. When we execute a SPARQL query, such as that shown in Figure 3, we get minimum Statement Templates that define only property constraints.

```
SELECT DISTINCT ?p
WHERE {
  ?s ?p ?o .
  ?s rdf:type foaf:Group .
  ?s rdf:type foaf:Organization .
  FILTER NOT EXISTS {
    ?s rdf:type ?type .
    FILTER(?type != foaf:Group)
    FILTER(?type != foaf:Organization)
  }
}
```

FIG. 3. A SPARQL query for extracting properties which instances of foaf:Group ∩ foaf:Organization have

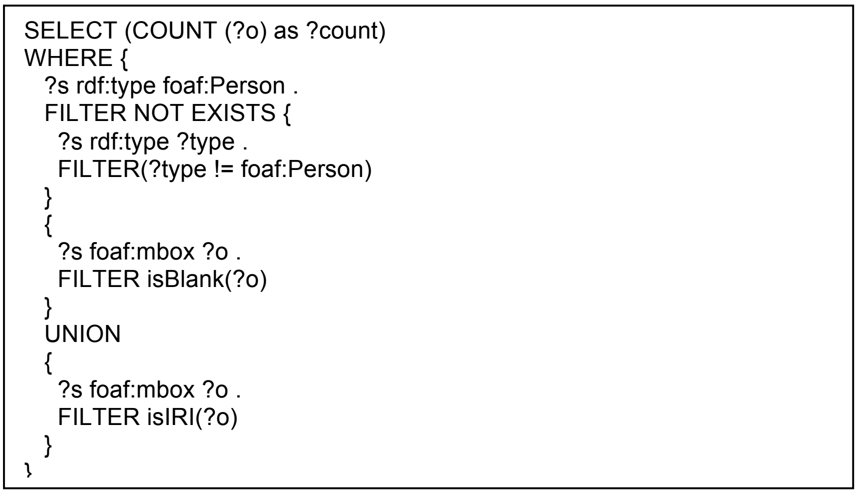
We estimate value constraints in the third step. After we get metadata values using classes of resources and a property, we classify those values into "literal", "non-literal" and "mix". To estimate value constraints, we count the number of the three metadata values below.

A) The number of all metadata values,

B) The number of literal metadata values, and

C) The number of non-literal metadata values.

When A = B, we define value constraints as "literal". If A > B and A > C, we define value constraints as "mix". For extracting B and C, we use *isLiteral*, *isIRI* and *isBlank*, which are SPARQL functions that are shown in a SPARQL query in Figure 4.



```
SELECT (COUNT (?o) as ?count)
WHERE {
  ?s rdf:type foaf:Person .
  FILTER NOT EXISTS {
    ?s rdf:type ?type .
    FILTER(?type != foaf:Person)
  }
  {
    ?s foaf:mbox ?o .
    FILTER isBlank(?o)
  }
  UNION
  {
    ?s foaf:mbox ?o .
    FILTER isIRI(?o)
  }
}
```

FIG. 4. A SPARQL query for extracting the number of non-literal metadata values for foaf:Person and foaf:mbox

In the final step, we extract the constraints of literal and non-literal metadata values such as class memberships of non-literal resources, base URIs, language tags and datatypes of literal metadata values. This process is executed based on the result of step 3. If the metadata type value is "non-literal", we extract resource classes and the base URI of metadata values by analyzing all objects data pulled back and load to the RDF database. When we found metadata with the value "literal", we defined datatype and language tags.

4. Evaluation

We implemented a system to extract DSPs using our approach. To evaluate our system and approach, we extract DSPs from 10 datasets and verify those DSPs. We used 10 LOD datasets that are published as RDF files on the DataHub[2]. It is difficult to extract metadata schema manually, so to evaluate our method for large datasets, we chose datasets that could be accessed on the Web and were the top 10 largest in file size at the time of access. In this evaluation, we confirm only precision by comparing constraints which are extracted using our approach and a manual method, and also comparing extracted constraints and actual datasets.

First, we compared structural constraints defined by DSPs, which were extracted by our approach and a manual method. Using this comparison, we attempted to confirm if the system we implemented is running correctly based on our proposed method. A person who executes a manual method has knowledge and experience of designing metadata schema, but may not have knowledge about the specific domain of each dataset (e.g., geography, statistics, etc.). In a manual method, the process of extracting a DSP is based on 5 steps that were shown in section 3. The difference of our approach and a manual method is the data size of RDF files. For extracting a DSP from a dataset, our approach used entire RDF files belonging to that dataset, whereas the manual method used the top 200 lines from each RDF file.

Table 3 shows the number of Description Templates and Statement Templates that were extracted using our approach and a manual method. We confirmed that all of structural constraints extracted manually were included in the structural constraints extracted by our approach. The constraints that we compared are shown in section 3. We also confirmed the constraints which were extracted by our approach are not contradictory to actual datasets. There are, however, differences between numbers of templates that were extracted by our approach and a manual method. One reason is because the amount of data that was used to manually extract was smaller than our approach. Another reason is that some resources have multiple RDF types,

[2] http://datahub.io/

Table 3: The number of Description Templates and Statement Templates that were extracted by our approach and a manual method

Dataset ID in the DataHub	Description Templates		Statement Templates	
	our approach	manual method	our approach	manual method
nytimes	1	1	13	9
colinda	2	1	15	7
mondial	19	4	107	31
eurostat-rdf	9	2	75	8
linked-open-vocabularies-lov	9	4	63	15
farmers-markets-geographic-data-united-states	33	4	164	18
msc	6	1	39	4
nuts-geovocab	4	3	15	11
osm-semantic-network	3	3	44	22
parole-simple-out	168	2	669	7

and those class memberships are difference each other. For these cases, we created a Description Template for each class membership, so that most Description Templates have only a few resources. For example, we extracted 168 Description Templates from *parole-simple-out*, but 106 Description Templates have less than 10 resources. We discuss this problem in section 5.

After we confirmed our system is running correctly, we checked constraints that were extracted by our method but weren't extracted by the manual method. As a result of comparing those constraints and original datasets, those constraints were not contradictory to datasets. Finally, we looked into parts of each dataset in order to find constraints which were not extracted by our method.

In the above procedure, we confirmed that it is possible to extract most structural constraints, which described in section 3, using our approach. However, there are constraints which we could not extract using our approach. We discuss whether or not the constraints that we extracted are useful to understand existing metadata structures in the next section.

5. Discussion

We could not extract structural constraints of resources which do not have *rdf:type* using our approach. For example, *nuts-geovocab,* for describing geographical metadata, includes RDF Collections in order to describe the exterior of geospatial objects with multiple coordinates. Figure 5 shows metadata instances from *nuts-geovocab*. There are more than two coordinates for describing the exterior of the resource *"http://nuts.geovocab.org/id/AT111_geometry"*. Those coordinates are described using non-typed blank nodes which are connected with *rdf:first* and *rdf:rest*. This meant that we could not extract the Description Templates for resources that describe coordinates. When we guess the classes of each resource using existing metadata schema and definitions about metadata vocabularies which include *rdfs:domain* or *rdfs:range*, we can extract more Description Templates.

There are other issues that need to be solved in order to improve our approach. In this evaluation, we could extract a large number of Description Profiles from *farmers-markets-geographic-data-united-states* and *parole-simple-out*. We proceeded to check their Description Templates and Statement Templates. As a result, in some cases, we could merge the Description Templates into other templates. For example, *farmers-markets-geographic-data-united-states*, there are the following two class memberships, Class membership defined in Description Template A

- http://logd.tw.rpi.edu/source/data-gov/vocab/Dataset (logd:Dataset)
- http://purl.org/twc/vocab/conversion/Dataset (conversion:Dataset)
- http://purl.org/twc/vocab/conversion/MetaDataset (conversion:MetaDataset)
- http://rdfs.org/ns/void#Dataset (void:Dataset)

Class membership defined in Description Template B

- http://logd.tw.rpi.edu/source/data-gov/vocab/Dataset (logd:Dataset)
- http://purl.org/twc/vocab/conversion/Dataset (conversion:Dataset)
- http://purl.org/twc/vocab/conversion/SameAsDataset (conversion:SameDataset)
- http://rdfs.org/ns/void#Dataset (void:Dataset)

Description Template A and B have differences in the two classes conversion:MetaDataset and conversion:SameDataset. Both Description Templates have 8 Statement Templates, and those Statement Templates are similar. If there are a large number of Description Templates, metadata schema designers cannot easily understand the structural constraints of the dataset. In that case, we should define one Description Template for resources which are instance of (logd:Dataset ∩ conversion:Dataset ∩ void:Dataset).

We believe that we are unable to extract DSPs correctly if there are resources that have multiple roles in the datasets. We have created and published *Aozora Bunko LOD*[3] which is a dataset including bibliographies based on *Aozora Bunko*[4]. *Aozora Bunko* is a Japanese digital library that publishes digitized books. The bibliographies, which are published on *Aozora Bunko*, have some resources about persons, such as "creator", "translator" and "reviser". We described person as a instance of *aozora:Person*. However, instances of *aozora:Person* have different roles in that dataset as mentioned above. In that case, we can only extract one Description Template about *aozora:Person*, and in the Description Template, the metadata attributes for the persons with different roles are mixed. There are two approaches to resolve this problem. One is by

```
<geometry:Polygon
  xmlns:geometry="http://geovocab.org/geometry#"
  xmlns:wgs84pos=" http://www.w3.org/2003/01/geo/wgs84_pos#"
  rdf:about="http://nuts.geovocab.org/id/AT111_geometry">
  <geometry:exterior>
    <geometry:LinearRing>
      <geometry:posList>
        <rdf:Description>
          <rdf:first>
            <rdf:Description>
              <wgs84pos:lat>47.35300025</wgs84:lat>
              <wgs84pos:long>16.435400050000055</wgs84:long>
            </rdf:Description>
          </rdf:first>
          <rdf:rest>
            <rdf:Description>
              <rdf:first>
                <rdf:Description>
                  <wgs84:lat>47.455132750000018</wgs84:lat>
                  <wgs84:long>16.281081050000068</ns48:long>
```

FIG. 5. An example of resource which are described using non-typed resources

[3] http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/

[4] http://www.aozora.gr.jp/

adding different classes for each type of person in the original datasets. Because it is required to change source data, this approach is not practical. The other is extracting a Description Template for each pair of a class membership and a property that has an instance of that class membership as a range. For example, if there are metadata instances which figure 6 shows, we should extract two Description Templates for *aozora:Person* as *dc:creator* and *aozora:Person* as *dc:translator*.

```
<book_A> dc:creator <person_X> ;
            dc:translator <person_Y> .

<person_X> rdf:type aozora:Person .
<person_Y> rdf:type aozora:Person .
```

FIG. 6. An examples of resources which are both instance of aozora:Person, and have different roles "dc:creator" and "dc:translator"

6. Conclusion

In this paper, we have proposed a method for extracting the structural constraints of LOD datasets using metadata instances and existing schema. Metadata schema about existing datasets are important for metadata schema designers to create a new interoperable schema with a low cost. However, because creating formal metadata schema is costly, there are few schema about existing LOD datasets on the web. We aim to extract metadata schema automatically, especially the structural constraints of metadata records, in order to add metadata schema to metadata schema registries.

To evaluate our approach, we compared the number of structural constraints which were extracted by our approach and manually with 10 datasets in the DataHub. That evaluation showed that our approach could extract all the structural constraints which could be extracted manually. We also compared metadata instances and structural constraints which are extracted using our approach. As a result, it has become clear that there are three issues to be solved when extracting structural constraints using our approach. One is the need to improve our method for extracting Description Templates of resources which have no *rdf:type*. The second issue is that we need to merge Description Templates when the extracted templates are similar to other templates. The last issue is that we separate templates for resources, which have same classes, but have different roles in a dataset.

References

Chidlovskii. Boris (2002). Schema extraction from XML collections. Proceedings of the 2nd ACM/IEEE-CS joint conference of Digital libraries, 2002, 291-292.

Coyle, Karen and Thomas Baker. (2009). Guidelines for Dublin Core Application Profiles. Retrieved May 15, 2014, from http://dublincore.org/documents/2009/05/18/profile-guidelines/ .

Hillmann, I. Diane I, Stuart A. Sutton, Jon Phipps and Ryan Laundry. (2006). A Metadata registry from vocabularies up: The NSDL registry project. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006.

Konrath, Mathias, Thomas Gottron, Steffen Staab and Ansgar Scherp. (2012). SchemEX – Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. Journal of Web Semantics, 2012, vol. 16.

Nagamori, Mitsuharu, Masahide Kanzaki, Naohisa Torigoshi and Shigeo Sugimoto. (2011). Meta-Bridge: A Development of Metadata Information Infrastructure in Japan. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2011, 63-68.

Nilsson, Mikael. (2008). Description Set Profiles: A constraint language for Dublin Core Application Profiles. Retrieve May 15, 2014, from http://dublincore.org/documents/dc-dsp/ .

Nilsson, Mikael, Thomas Baker and Pete Johnston. (2008). The Singapore Framework form Dublin Core Application Profiles. Retrieve May 15, 2014, from http://dublincore.org/documents/2008/01/14/singapore-framework/ .

Nishide, Yoritsugu, Tsunagu Honma and Mitsuharu Nagamori. (2013). An Investigation of Japanese Open Data Schema and Links to Improve the Use of Datasets. Digital Library, 2014.

Infrastructure & Models—Part B

The *1:1 Principle* in the Age of Linked Data

Richard J. Urban
Florida State University
School of Information
rurban@fsu.edu

Abstract

This paper explores the origins of the *1:1 Principle* within Dublin Core Metadata Initiative (DCMI). It finds that the need for the *1:1 Principle* emerged from prior work among cultural heritage professionals responsible for describing reproductions and surrogate resources using traditional cataloging methods. As the solutions to these problems encountered new ways to model semantic data that emerged outside of libraries, archives, and museums, tensions arose within DCMI community. This paper aims to fill the gaps in our understanding of the *1:1 Principle* by outlining the conceptual foundations that led to its inclusion in DCMI documentation, how the *Principle* has been (mis)understood in practice, how violations of the *Principle* have been operationalized, and how the fundamental issues raised by the *Principle* continue to challenge us today. This discussion situates the *1:1 Principle* within larger discussions about cataloging practice and emerging Linked Data approaches.

Keywords: 1:1 Principle, RDF, Abstract Model,

1. Introduction

> In general, Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand in for one another. For instance, a jpeg image of the *Mona Lisa* has much in common with the original painting, but it is not the same as the painting. As such the digital image should be described as itself, most likely with the creator of the digital image included as a Creator or Contributor, rather than just the painter of the original *Mona Lis*a. The relationship between the metadata for the original and the reproduction is part of the metadata description, and assists the user in determining whether he or she needs to go to the Louvre for the original, or whether his/her need can be met by a reproduction (Hillmann, 2003).

The Dublin Core Metadata Initiative (DCMI) *1:1 Principle* appears to offer a simple dictum: “metadata is about one, and only one, resource” (Powell, Nilsson, Naeve, Johnston, & Baker, 2007).[1] Yet despite its apparent simplicity, “one to one…is a many headed snake, and it has bitten us often over the years.” (Weibel, 2010). Metadata creators find the *Principle* confusing or, at best, routinely ignore it because it remains unsupported by digital library software and exchange protocols (Han, Cho, Cole, & Jackson, 2009; Hutt & Riley, 2005; S. J. Miller, 2010; Park & Childress, 2009; Park, 2009; Shreeves et al., 2005; Stvilia, et al., 2004; Urban, 2012). Although the specific definition provided in Hillmann’s (2003) Using Dublin Core (and the “one-to-one” label itself) has fallen out of favor, the fundamental questions embodied in the *Principle* continue to animate debates and discussions about the DCMI Abstract Model and DCMI’s relationship to the Resource Description Framework (RDF).

This paper aims to fill the gaps in our understanding of the *1:1 Principle* by outlining the conceptual foundations that led to its inclusion in DCMI documentation, how the *Principle* has been (mis)understood in practice, how violations of the *Principle* have been operationalized, and how the fundamental issues raised by the *Principle* continue to challenge us today. This

[1] For consistency, I use *1:1 Principle* except when variants are used in direct quotes. i.e. “one-to-one,” etc.

discussion situates the *1:1 Principle* within larger discussions about cataloging practice and semantic knowledge representations.

2. Background

While the specifics of the *1:1 Principle* are directly tied to the development of Dublin Core (DC), the general problem that it references — how to model the description of original resources and their associated reproductions or surrogates in various formats — is one that has plagued cataloging standards since reproductive technologies (such as photography, microfilm, and microfiche) became widely available in the mid-20th century. At the heart of these discussions are ontological distinctions among different kinds of bibliographic entities (e.g. multiple versions, electronic resources, non-book resources). But is also an account of how flat bibliographic records have struggled to represent the complex relationships among these entities. At the time that DC was being defined in the mid-1990s, many of the key stakeholders in its development had already been wrestling with these issues for more than a decade.

2.1. Describing Reproductions, Multiple Versions, and Electronic Resources

From the earliest cataloging guidelines, concerns about representing "reproductions" of bibliographic materials complicated emerging descriptive standards. As libraries began collecting an increasing number of different reproductive media (microfilms and microfiche), or multiple versions of the same work (i.e. a musical recording released simultaneously on vinyl, cassette, and/or compact disc), the problems began to multiply (Graham, 1992; Knowlton, 2009). Simonton's report (1962), commissioned by the Association of Research Libraries (ARL), defined two solutions to the problem that serve as the foundations for current practice:

- The *Facsimile Theory* privileged the intellectual content of an item by making the "original" resource the focus of the record representing a reproduction. Following the long-standing practice of dash entries, a description of the reproduction itself would be included as a note.
- The *Edition Theory* required a record to represent the physical features of the reproduction, using a note to provide a description of the "original" resource.

The first edition of the *Anglo-American Cataloging Rules (AACR1)* used the facsimile theory and dashed entries to continue a common practice. However, *AACR2's* cardinal principle required a shift in cataloging rules towards an edition theory (item-at-hand) perspective (Graham, 1992).[2]

This shift was not welcomed by the cataloging community who "assailed [it] as 'an obsession with principle to the exclusion of common sense'" (Graham, 1992). Most vocal in their opposition to the rule change were libraries and information centers that dealt in large numbers of "reproduction" records, such as the Library of Congress (LOC), the National Library of Medicine (NLM), and academic libraries participating in the NEH-funded U.S. Newspaper Program (USNP). In response, the LOC issued a rule interpretation upholding a facsimile theory approach (Graham, 1992; Library of Congress, 2010). While some bibliographic services, such as the Research Libraries Group (RLG) RLIN, adapted to these rule interpretations, many cataloging services could not take full advantage of them, leaving "a fractured set of approaches" in place (Jones, 1997). Following the precedent set with microfilm reproductions, the Library of Congress applied the same rule interpretation to the digitization of its photography collections (Arms, 1999). "The records describe the intellectual expression and the original form of the material and provide a link to the corresponding digital reproductions" (Library of Congress, 2010).

Many of the arguments about which theory should be used center around user needs and the functions of information retrieval systems. For example, an advantage of the facsimile theory is that it allowed records about originals and reproductions to co-locate in the catalog, thereby

[2] "The starting point for description is the physical form of the item at hand, not the original or any previous form in which the work has been published" (American Library Association, et al., 1988).

saving the time of the user. The facsimile theory also had economic advantages. Under an edition theory approach (*AACR2*), a cataloger had to "start over" to create a new record for the reproduction. The facsimile theory (*AACR1/LOC 1.11*) allowed catalogers to quickly clone an existing description and append a reproduction note, saving significant costs. (Graham, 1992).

2.2. Beyond the Book: The Description of Art, Visual Resources, and Archival Materials

At the same time that cataloging standards struggled with reproductions a parallel conversation was taking place about the representation of surrogates for non-book visual materials, such as artworks, photography, and archival materials. Members of this community drew careful distinctions between a reproduction that fully represented an original object and surrogates which merely stood-in for the object, i.e. a photograph of a 3-dimensional sculpture does not reproduce the sculpture, but does allow us to represent it in an information system. This community included professionals responsible for managing visual resource collections (art and architectural slide collections) and museum collections (the Getty's Art History Information Program, later the Getty Information Institute – GII) (Fink, 1999; McRae & White, 1998). Until the advent of centralized online catalogs, the distinction between originals and surrogates was handled by establishing physically separate card catalogs. However, in a MARC-based catalog what kind of resource a record represented was less clear. In order to make this more explicit, the MARC Visual Materials (MARC-VM) and Archival Materials Control (MARC-AMC) formats introduced new control fields that made the "type of record" explicit (Dooley & Zinham, 1990; Evans & Will, 1988). In discussing the need for these new features, we see examples that would later be revisited to illustrate the need for the *1:1 Principle*:

> The [*Art and Architecture Thesurus*] considers reproductions of works of art to be surrogates for original works and will recommend that they be indexed in a similar fashion. For example, PAINTING (655) would be used to describe both Leonardo's *Mona Lisa* and a slide reproduction; SLIDE (655) would also be used in the latter case. This holds serious implications for effective retrieval….In an integrated database containing both of these media, searchers interested only in examples of actual paintings might have to learn to exclude slides, microfilm, and other reproduction media in their search queries to retrieve only records for original paintings. . . . One solution might be the addition of a "reproduction" facet to indexing strings for object surrogates so that they would be differentiated from "originals" in a browse display (Dooley & Zinham, 1990).

The ability to distinguish between descriptions of originals and surrogates in various analog and digital formats was a key component of emerging standards for describing information about artworks and museum objects. Both the Categories for the Description of Works of Art (CDWA) and the Visual Resource Association's VRACore included structures that enabled the separation of information about different kinds of resources (Baca, 2002; Harpring & Baca, 2009; Visual Resources Association & Whiteside, 1999).

2.3. A Principle is Born

When the DCMI began, it had an explicit goal to describe "document-like objects" (DLO) found on the World Wide Web (Weibel, 1995). The development of this new standard soon came to the attention of several organizations interested in developing online representations for their collections, including RLG, the Getty Information Institute (GII), and the UKOLN Arts and Humanities Data Service (AHDS) (Erway, 1996; Fink, 1999; P. Miller & Greenstein, 1997). Advocating for the needs of library, archive, and museum (LAM) collections, RLG argued that DC could be used to describe offline physical collections and that the definition of DLOs should extend to images (Erway, 1996). The *Guidelines for Extending the Use of Dublin Core Elements* grounded its recommendations for a "record type" indicator or element refinements on earlier work for reproduction/surrogate descriptions (Research Libraries Group, 1997a, Research Libraries Group, 1997b).

The RLG proposal became a central point of discussion at the 1997 DC-4 Workshop in Helsinki, Finland. Rather than adopt the proposed changes in the RLG *Guidelines,* workshop participants discussed the relationship between "logical clusters of metadata…that reference one, and only one, state of the information resource," which became the nucleus of the *1:1 Principle* (Bearman, 1999; Weibel & Hakala, 1998).

Following the Helsinki meeting, *1:1 Principle* issues emerged in several working groups (One-to-One, Relations, and Data Model). The discussions were frequently contentious debates between members in different camps. Cultural heritage professionals' concerns with the *1:1 Principle* primarily focused on the kinds of resources that could be described using DC. Drawing on their experiences with previous standardization efforts, this camp felt it necessary to provide guidance for different types of materials. However, there was a strong resistance to DCMI getting into the cataloging rules business, especially ones that needed to deal with complexities of different ontological kinds. The members of this group preferred to let Dublin Core remain a simple vocabulary for resource discovery. Acknowledging the concerns of cultural heritage professionals, the latter group argued that the kind of discrimination sought for cultural materials could be handled by more robust local standards (P. Miller & Greenstein, 1997). Furthermore, discussions on the dc-one2one listserv:

> . . . made absolutely clear that there is no consensus on what 1:1 really means in practice. In the end, people will describe what *they* want to describe, for their purposes and the purposes of their user community. That means they may describe a TIFF of an Ansel Adams photograph as having been created by Ansel Adams. Who's to say they're wrong? (Wendler, 1999)

By the end of 1999, discussion in the One-to-One group dwindled without having reached a clear consensus on the *Principle.* It was formally combined with other task groups into the DC-Architecture working group which attacked the problem from a different perspective.

Discussions in the Relation working group focused more on developing logical clusters of metadata that could be linked together. The discussions echoed concerns found in earlier MARC-based solutions to representing originals and reproductions. In particular, there were concerns that separating descriptions into distinct records could result in a loss of information when shared outside of an application. The suggestion of separate records also raised concerns about how to display them to users, with a sense that independent representations of originals and reproductions would make the task harder. Proponents of "keeping Dublin Core simple" suggested that atomic statements about resources enabled better discovery of resources without the additional complexity of resource type-based models. Instead, statements about resources could be dynamically organized into logical packages for particular uses such as retrieval or display for a user (Lagoze, 1997, 2001a).

3. From Principle to Abstract Model

Thus far, the story of the *1:1 Principle* has been about cataloging practices in a cultural heritage community concerned with ontological distinctions and relationships among resources. The introduction of these concerns into the development of DC metadata brought these practices into contact with fundamentally different theories of description that emerged from formal knowledge representation (KR) approaches. KR semantics were not merely concerned with fixing the meaning of individual vocabulary terms, but how descriptions could consistently refer to described resources (Urban, 2012).

This was of little concerned when Dublin Core was created as embedded metadata within a document-like object, such as a HTML page. In this case the metadata described the resource that it was embedded within. A desire to describe non-textual resources meant developing a standalone Standard Generalized Markup Language (SGML) syntax that would provide "explicit semantics of each Dublin Core element"; however, "discrete packages of metadata cannot be identified and the semantics of repeated elements are not specified" (Burnard, Miller, Quin, & Sperberg-McQueen, 1996). These conversations resulted in the emergence of the Warwick Framework that would allow for the creation and exchange of metadata containers (Dempsey & Weibel, 1996; Lagoze, 1996). A package might include DC metadata, or metadata in other formats.

The Warwick Framework became one of several alternative metadata proposals submitted to the World Wide Web Consortium (W3C) in order to address laws aimed at filtering adult content on the Web. Among the others were the Platform for Internet Content Selection (PICS), Microsoft's XML Web Collections (XMLWC), and Apple's Meta Content Framework (MCF). Rather than developing each of these recommendations separately, the W3C rolled them together into a new initiative known as the Resource Description Framework (RDF) (E. Miller, 1998).

As a model for expressing a formal semantics for metadata, RDF owes a great deal to earlier artificial intelligence and knowledge representation research that took place before the advent of the World Wide Web (Halpin, 2004). In addition to fixing the meaning of properties used to describe resources, researchers in this area quickly realized that referent tracking was essential to the development of computational reasoning (Lenat & Guha, 1990). Guha would add features originally developed for the Cyc project to MCF and ultimately to RDF (Halpin, 2004). In the context of the RDF model, the relationship between a metadata statement and a resource is established through the consistent assignment of a URI (Berners-Lee, 2002; Hayes, 2004). In theory, if all the objects of description are supplied with a URI, statements about those resources will naturally organize themselves around these identifiers, fulfilling the main objectives of the *1:1 Principle*.

The development of RDF and eXtensible Markup Language (XML) specifications encouraged DCMI to begin work on a more formal data model for Dublin Core (Baker, 2012; Weibel & Hakala, 1998; Weibel, 2010). Initially, this work expressed DC descriptions as a variant of RDF. However, within the implementer community, there was a great deal of initial resistance to RDF in favor of simpler "plain" XML representations. This was due in part to a lack of practice and software tools that could understand RDF, and to fundamental misunderstandings within the Dublin Core implementer community that saw RDF as an overly complex XML syntax (Baker & Johnston, 2011; Baker, 2012). Because the XML serialization of RDF represented a graph structure, it was also less human-readable than a document-like encoding of element/value pairs. Resistance to RDF also came from the Open Archives Initiative (OAI) community, which was developing a protocol for exchanging "packages" of metadata along the lines of the Warwick Framework. "It may be that the vast majority of data providers don't need (or even understand) RDF and are mainly interested in exposing metadata as simple attribute-value pairs or simple trees for which XML is perfectly appropriate" (Lagoze, 2001b). In order to conform to the simple DC and to provide a low barrier to use (i.e., by using well-supported technologies), OAI-PMH initially required a minimal DC XML schema (later versions of OAI-PMH referenced official DCMI XML syntax recommendations) (Lagoze, Van de Sompel, Nelson, & Warner, 2008). As a container architecture, OAI-PMH left the aboutness of a record to the enclosed metadata specification.

The intersection of XML and RDF models for DC metadata created some inherent tensions. Although DCMI developed an implicit grammar for statements, it was intentionally scruffy in order to accommodate the broad diversity emerging on the Web (Baker, 2000, 2012; Johnston, 2006). Addressing calls for more guidance, DCMI released official recommendations for encoding Dublin Core in XML and RDF that included rudimentary definitions of an abstract model. This initial model specified a one-to-one relationship between a record and a resource at the same time recognizing that "there is no formal linkage between a simple DC record and the resource being described. Such a linkage may be made by encoding the URI of the resource as the value of the DC Identifier element, however this is not mandatory" (Powell &

Johnston, 2002). Because of implementation confusions about this early model, a more formal recommendation was published as the DCMI Abstract Model (DCAM) (Powell, Nilsson, Naeve, Johnston, & Baker, 2005). Although DCAM borrowed some concepts from RDF, "DCAM was meant to provide a basis for guidelines that would allow metadata records to be encoded using XML, HTML, and in principle, any concrete implementation syntax…" (Baker, 2012, p. 121). Although DCAM enabled syntaxes to include "slots" for URIs to reference a resource, it also continued to support *1:1 Principle* concepts:

> The abstract model described above indicates that each DCMI metadata description describes one, and only one, resource. This is commonly referred to as the one-to-one principle…However, real-world metadata applications tend to be based on loosely grouped sets of descriptions (where the described resources are typically related in some way), known here as description sets. For example, a description set might comprise descriptions of both a painting and the artist…(Powell et al., 2005)

Unfortunately, DCAM failed to achieve widespread adoption within the Dublin Core implementer community, especially among LAMs that are the focus of this discussion. Instead of resolving the tensions between RDF and XML approaches, the DCAM "fell between two stools," leaving neither group invested in applying it to their data (Baker & Johnston, 2011).

4. *1:1 Principle* Violations and Metadata Quality

Because one of the fundamental objectives of Dublin Core is to enable to exchange of interoperable metadata, studying metadata quality has been an important activity. Among studies that examine DC metadata for cultural heritage resources, failure to comply with the *1:1 Principle* has been identified as cause for many quality problems (Han et al., 2009; Hutt & Riley, 2005; S. J. Miller, 2010; Park & Childress, 2009; Park, 2005; Shreeves et al., 2005; Stvilia et al., 2004).

For Shreeves, et al (2005), the *1:1 Principle* is related to the internal cohesiveness of a metadata record and the degree to which it represents related resources. In examining an aggregation of cultural heritage metadata, they found that "…no collection maintained a consistent one-to-one mapping between the metadata and a single resource…" Within an individual collection, "between 57% and 100% of records in their sample included properties for both physical and digital manifestations of a resource" (Shreeves et al., 2005). These findings were later confirmed by Hutt & Riley (2005), Han, et al (2009) and again by S. J. Miller (2010).

S.J. Miller (2010) notes that *1:1 Principle* problems result from "database and user interface systems [that] do not have the capacity to adequately link separate records and to display them together in a clear and meaningful way for end users." Systems, such as CONTENTdm, base their primary information models around digital assets, making it difficult to independently represent non-digital source resources (Han et al., 2009). These systems also enable metadata creators to add specialized, locally defined metadata elements on a collection-by-collection or project-by-project basis. The ease with which these systems allow the addition of new properties encourages ad-hoc modeling optimized for display in one local context, rather than more formal and rigorous methods of modeling on at Web scale.

4.1. Limitations of Violations

In light of the debates that brought the *1:1 Principle* into existence, it is necessary to question many of the assumptions that have gone into quality studies. First, the studies themselves demonstrate that the *1:1 Principle* was not necessarily a concern among metadata creators. Instead, conforming to cataloging rules for reproductions and/or surrogate resources provided the context for descriptions. Regardless of whether an record uses facsimile (*AACR1*) or edition (*AACR2*) theory approaches, MARC inherently describes more than one resource. While local practices for Dublin Core may not alter the definition of DC terms, they implicitly changed the referent to a different resource (i.e. the prevalence of date.original, date.digital). The adoption of these rules in association with Dublin Core, particularly within the library community, is often justified by user convenience and economics (Cronin, 2008; S. J. Miller, 2010).

Secondly, most of these studies use a "record" as the unit of analysis for assessing metadata quality, especially the set of DC elements provided by an OAI-PMH DC record. As noted above, oai_dc is based on a 2002 XML schema recommendation that pre-dates DCAM (Lagoze, Van de Sompel, Nelson, & Warner, 2002; Lagoze et al., 2008). Neither the OAI-PMH container architecture nor this Dublin Core schema enable DCAM-like description sets that would comply with the *1:1 Principle*. These problems are further compounded by the limitations of data representations within commonly used digital repository systems like ContentDM (Han, et al., 2009, S. J. Miller, 2010).

Furthermore, these studies are only able to detect a limited set of *1:1 Principle* violations. Most operationalize violations of the *1:1 Principle* through a conjunction of oai_dc statements (i.e., the resource hasFormat "image/jpeg" AND hasFormat "oil on board"). Although the informal definition of a *1:1 Principle* licenses such an assumption, it is not supported formally by the XML semantics or the DCAM. The detection of *1:1 Principle* violations has hinged on format and date elements that supply ontological absurdities. Being aware that metadata represents cultural heritage resources heightens our awareness of incoherent format statements that describe the properties of both physical and digital resources. In a heuristic evaluation of metadata records, qualitative researchers bring a great deal of background knowledge to their assessments. They may intuitively understand that terms like image/jpeg and glass plate negative are properties that are unlikely to be shared by the same resource. They also may understand that JPEGs are the kinds of the resource that "reproduce" something like a glass plate negative, but rarely will glass plate negatives "reproduce" a JPEG. They understand that JPEGs are the kind of resource that can be associated with "2008" and are not resources that could have been created in "1901." These kinds of inferences are difficult to automate even when using robust taxonomies because they require integrating and aligning knowledge from across multiple sources (for example, *AAT* knows little about specific file formats described in a resource such as the Unified Digital Format Registry (UDFR)). Even accepting these limitations, these automated approaches fail to identify violations when DC records appear to be internally coherent. For example a DC description of a microfilm that merely uses a URL to link to a digitized version of the resource.

5. Would RDF save us from *1:1 Principle* Violations*?*

The studies discussed above all took OAI-PMH XML as their focus, leaving an important question unanswered: Would an RDF-based approach save us from rampant violations of the *1:1 Principle*? Debates from within the Semantic Web/Linked Data community suggest that RDF alone does not solve the problems inherent in the *1:1 Principle* but rather shifts the burden onto URIs. Known as the Semantic Web Identity Crisis or http range-14 problem, the debates on this issue closely parallel *1:1 Principle* problems (Halpin, 2011; Hayes & Halpin, 2008). At the heart of the problem is the question of whether a URI can refer to both an information object that describes an entity (i.e., a surrogate representation) and the entity being described. Hayes and Halpin (2008) provide the example of a URI that may refer to the Eiffel Tower itself (the structure in Paris designed by Gustave Eiffel) and a photograph of the Eiffel Tower (or equally, a set of RDF statements about the Eiffel Tower). According to Hayes & Halpin, what a URI refers to may be specified by the formal interpretation associated with it. In one interpretation, the URI may refer to the surrogate representation (the photo); in another, it may refer to the entity the surrogate stands for (the Eiffel Tower itself). In contrast, Berners-Lee (2002) argues that URIs refer to one, and only one, resource, as determined by the agent responsible for "minting" the URI (in part through the authority bestowed by the owner of a domain name). To date, World Wide Web Consortium (W3C) recommendations support Berners-Lee's approach (Sauermann & Cyganiak, 2008). However in a study of available Linked Data, Halpin, et al (2010) found that the same Linked Data URI was being used to refer to distinct entities in different contexts (for example, the city of Paris as a political entity vs. Paris as a geographic location). Within the present metadata quality literature, the question of whether a URI successfully refers to the described resource is left unmeasured, especially for the use of URIs that do not provide access to offline resources, but may successfully refer to them. While identifiers found in OAI-PMH records had a high degree of uniqueness, this does not entail that any identifier refers uniquely to one, and only one, resource. This suggests that another kind of *1:1 Principle* violation may occur if a URI is used to refer to more than one resource (Stvilia et al., 2004; Stvilia & Gasser, 2008).

6. Conclusion

The developers of Dublin Core intended it to be a simple vocabulary that could be broadly applied to emerging Internet resources. The introduction of cultural heritage material introduced more complex kinds of relationships between online and offline resources or "originals" and "reproductions." Faced with this problem, the cultural heritage community proposed solutions based on many years of practice using document surrogates in information retrieval systems. However, users of traditional cataloging systems also struggled with defining best practices for describing reproductions and multiple versions. Conflicting interpretations meant that document surrogates could appear in two forms based on the object of description (i.e., facsimile/edition theory approaches). Within the DCMI, these developments in descriptive cataloging encountered new approaches to representing descriptions as "metadata." While emerging technologies such as XML enabled the creation of document-like data models, the development of DC was also influenced by more formal modeling techniques, such as RDF, that required a one-to-one relationship between entities and their descriptions. Because this requirement conflicted with the cultural heritage community's recommendations for handling reproductions, it was necessary to articulate it in DCMI documentation as the *1:1 Principle*. However, these recommendations failed to overcome the limitations the cultural heritage community's pragmatic understanding of the relationship between descriptions and resources. While the limitations of systems for storing and exchanging DC metadata are implicated in the prevalence of *1:1 Principle* problems, there also seemed to be little desire from within the community for more formal representation models, such as RDF. However, it is important to recognize that RDF, in and of itself, is insufficient to solve fundamental identity issues embodied by the *1:1 Principle.* The more recent development of complex bibliographic models, such as Functional Requirements for Bibliographic Records (FRBR), and their implementation as Linked Data, suggest opportunities to reformulate our ability to detect whether a description is about "one and only one resource."

References

American Library Association, Australian Committee on Cataloguing, Canadian Committee on Cataloguing, British Library, & Library of Congress. (1988). *Anglo-American Cataloging Rules*. (M. Gorman & P. W. Winkler, Eds.) (2nd Edition revised.). Chicago: American Library Association.

Arms, C. R. (1999). Getting the picture: Observations from the library of congress on providing online access to pictorial images. *Library Trends*, *48*(2), 379–409.

Baca, M. (Ed.). (2002). *Introduction to Art Image Access: Issues, Tools, Standards, Strategies*. Los Angeles: Getty Research Institute. Retrieved from http://www.getty.edu/research/conducting_research/standards/intro_aia/

Baker, T. (2000). A grammar of Dublin Core. *D-Lib Magazine*, *6*(10). Retrieved from http://www.dlib.org/dlib/october00/baker/10baker.html

Baker, T. (2012). Libraries, languages of description, and linked data: a Dublin Core perspective. *Library Hi Tech*, *30*(1), 116–133.

Baker, T., & Johnston, P. (2011, May 13). Review of DCMI Abstract Model. Dublin Core Metadata Initiative. Retrieved from http://wiki.dublincore.org/index.php/Review_of_DCMI_Abstract_Model

Bearman, D. (1999, January). A common model to support interoperable metadata: Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities. *D-Lib Magazine*, *5*(1). Retrieved from http://www.dlib.org/dlib/january99/bearman/01bearman.html

Berners-Lee, T. (2002, July 27). What do URIs identify? W3C. Retrieved from http://www.w3.org/DesignIssues/HTTP-URI.html

Burnard, L., Miller, E., Quin, L., & Sperberg-McQueen, C. M. (1996, April 1). A syntax for Dublin Core Metadata. Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/workshops/dc2/report-19960401.shtml

Cronin, C. (2008). Metadata provision and standards development at the Collaborative Digitization Program (CDP): A history. *First Monday*, *13*(5). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2085/1957

Dempsey, L., & Weibel, S. L. (1996). The Warwick Metadata Workshop: A framework for the deployment of resource description. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/july96/07weibel.html

Dooley, J. M., & Zinham, H. (1990). The object as "subject": Providing access to genres, forms of materials, and physical characteristics. In P. Molholt & T. Petersen (Eds.), *Beyond the Book: Extending MARC for Subject Access* (pp. 43–80). Boston, MA: G.K. Hall & Co.

Erway, R. (1996). Digital initiatives of the Research Libraries Group. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/december96/rlg/12erway.html

Evans, L. J., & Will, M. O. (1988). *MARC for Archival Visual Materials*. Chicago: Chicago Historical Society.

Fink, E. (1999). The Getty Information Institute: A retrospective. *D-Lib Magazine*, *5*(3). Retrieved from http://www.dlib.org/dlib/march99/fink/03fink.html

Graham, C. (1992). Microform reproductions and multiple versions. *The Serials Librarian*, *22*(1), 213–234. doi:10.1300/J123v22n01_14

Halpin, H. (2004). The Semantic Web: The origins of artificial intelligence redux. Presented at the Third International Workshop on the History and Philosophy of Logic, Mathematics, and Computation (HPLMC-04 2005), Donostia San Sebastian, Spain. Retrieved from http://www.ibiblio.org/hhalpin/homepage/publications/airedux.pdf

Halpin, H. (2011). Sense and reference on the Web. *Minds and Machines*, *21*(2), 153–178. doi:10.1007/s11023-011-9230-6

Halpin, H., Hayes, P. J., McCusker, J. P., McGuinness, D. L., & Thompson, H. S. (2010). When owl:sameAs isn't the same: An analysis of identity in Linked Data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, … B. Glimm (Eds.), *The Semantic Web – ISWC 2010* (Vol. 6496, pp. 305–320). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://www.springerlink.com/content/v24433851k747864/

Han, M.-J., Cho, C., Cole, T., & Jackson, A. (2009). Metadata for special collections in CONTENTdm: How to improve interoperability of unique fields through OAI-PMH. *Journal of Library Metadata*, *9*(3), 213–238. doi:10.1080/19386380903405124

Harpring, P., & Baca, M. (Eds.). (2009). Categories for the Description of Works of Art. J. Paul Getty Trust. Retrieved from http://www.getty.edu/research/conducting_research/standards/cdwa/

Hayes, P. J. (2004). RDF Semantics. W3C. Retrieved from http://www.w3.org/TR/2004/REC-rdf-mt-20040210/

Hayes, P. J., & Halpin, H. (2008). In defense of ambiguity. *International Journal on Semantic Web and Information Systems*, *4*(2), 1–18.

Hillmann, D. (2003, August 26). Using Dublin Core. Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/documents/2003/08/26/usageguide/

Hutt, A., & Riley, J. (2005). Semantics and syntax of Dublin Core usage in Open Archives Initiative data providers of cultural heritage materials. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries* (p. 270).

Johnston, P. (2006, November 28). Why an abstract model for Dublin Core metadata? *eFoundations*. Retrieved from http://efoundations.typepad.com/efoundations/2006/11/why_an_abstract.html

Jones, E. (1997). Multiple Versions Revisited. *The Serials Librarian*, *32*(1), 177–198. doi:10.1300/J123v32n01_14

Knowlton, S. A. (2009). How the current draft of RDA addresses the cataloging of reproductions, facsimiles, and microforms. *Library Resources and Technical Services*, *53*(3), 159–165.

Lagoze, C. (1996). The Warwick Framework: A container architecture for diverse sets of metadata. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

Lagoze, C. (1997). From static to dynamic surrogates: Resource discovery in the digital age. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/june97/06lagoze.html

Lagoze, C. (2001a). Keeping Dublin Core simple: Cross-domain discovery or resource description? *D-Lib Magazine*, *7*(1). Retrieved from http://dlib.anu.edu.au/dlib/january01/lagoze/01lagoze.html

Lagoze, C. (2001b, May 17). RE: RDF, OAI, and application within libraries.

Lagoze, C., & Van de Sompel, H. (2008, October 17). ORE Specification - Abstract Data Model. Open Archives Initiative. Retrieved from http://www.openarchives.org/ore/1.0/datamodel#Foundations

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). Implementation guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative. Retrieved from http://www.openarchives.org/OAI/2.0/guidelines.htm

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2008). Open Archives Initiative Protocol for Metadata Harvesting. (OAI Executive & OAI Technical Committee, Eds.). Open Archives Initiative. Retrieved from http://www.openarchives.org/OAI/openarchivesprotocol.html

Lenat, D. B., & Guha, R. V. (1990). *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley Publishing Company.

Library of Congress. (2010). 1.11A Facsimiles, Photocopies, and Other Reproductions. Library of Congress. Retrieved from http://www.loc.gov/cds/PDFdownloads/lcri/LCRI_2010-03.pdf

McRae, L., & White, L. S. (Eds.). (1998). *ArtMARC Sourcebook: Cataloging Art, Architecture and their Visual Images*. Chicago, IL: American Library Association.

Miller, E. (1998). An introduction to the Resource Description Framework. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/may98/miller/05miller.html

Miller, P., & Greenstein, D. (Eds.). (1997). *Discovering Online Resources Across the Humanities: A Practical Implementation of Dublin Core*. London: UKOLN.

Miller, S. J. (2010). The One-to-One Principle: Challenges in Current Practice. *International Conference on Dublin Core and Metadata Applications*. Retrieved from http://dcpapers.dublincore.org/ojs/pubs/article/view/1043/992.

Park, J. (2005). Semantic interoperability across digital image collections: a pilot study on metadata mapping. *Lecture Notes in Computer Science*, *3237*, 621–630.

Park, J. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, *47*(3), 213–228. doi:10.1080/01639370902737240

Park, J., & Childress, E. (2009). Dublin Core metadata semantics: An analysis of the perspectives of information professionals. *Journal of Information Science*, *XX*(X), 1–13. doi:10.1177/0165551509337871

Powell, A., & Johnston, P. (2002, January 31). Guidelines for implementing Dublin Core in XML. UKOLN. Retrieved from http://www.ukoln.ac.uk/metadata/dcmi/dc-xml-guidelines/2002-01-31/#DCARCH

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2005, March 7). DCMI Abstract Model. Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/documents/2005/03/07/abstract-model/

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007). DCMI Abstract Model. Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/documents/abstract-model/

Research Libraries Group. (1997a). Guidelines for extending the use of Dublin Core Elements. Retrieved October 1, 2010, from http://www.oclc.org/research/activities/past/rlg/dcmetadata/guidelines.htm

Research Libraries Group. (1997b). Metadata Summit summary. Retrieved October 1, 2010, from http://www.oclc.org/research/activities/past/rlg/dcmetadata/summit.htm

Sauermann, L., & Cyganiak, R. (2008, November 3). Cool URIs for the Semantic Web. W3C. Retrieved from http://www.w3.org/TR/cooluris/

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is "quality" metadata "shareable" Metadata? The implications of local metadata practices for federated collections. In *Currents and convergence: navigating the rivers of change: proceedings of the Twelfth National Conference of the Association of College and Research Libraries April 7-10, 2005, Minneapolis, Minnesota* (p. 223).

Simonton, W. (1962). The bibliographic control of microforms. *Library Resources & Technical Services*, *6*(1), 29–40.

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. *Library and Information Science Research*, *30*(1), 67–74.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S. L., & Cole, T. W. (2004). Metadata quality for federated collections. In *Proceedings of ICIQ04-9th International Conference on Information Quality* (pp. 111–125).

Urban, R. J. (2012). *Principle paradigms: Revisiting the Dublin Core 1:1 Principle* (Dissertation). University of Illinois at Urbana-Champaign, Urbana, IL. Retrieved from http://hdl.handle.net/2142/31109

Visual Resources Association, & Whiteside, A. (1999, December 1). The core categories for visual resources - introduction. Retrieved September 26, 2010, from http://web.archive.org/web/20010306092716/www.gsd.harvard.edu/~staffaw3/vra/coreintro.htm

Weibel, S. L. (1995). Metadata: the foundations of resource description. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/July95/07weibel.html

Weibel, S. L. (2010). Dublin Core Metadata Initiative (DCMI): A personal history. In *Encyclopdia of Library and Information Sciences*.

Weibel, S. L., & Hakala, J. (1998, February). DC-5: The Helsinki Metadata Workshop; A Report on the workshop and subsequent developments. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/february98/02weibel.html

Wendler, R. (1999, April 14). Re: 1:1 debate: What's the goal? *dc-one2one*. Retrieved from http://dublincore.org/groups/one2one/

Towards Description Set Profiles for RDF using SPARQL as Intermediate Language

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

Kai Eckert
Research Group Data and Web Science
University of Mannheim, Germany
kai@informatik.uni-mannheim.de

Abstract

Description Set Profiles (DSP) are used to formulate constraints on valid data within a Dublin Core Application Profile. For RDF, SPARQL is generally seen as the method of choice to validate data according to certain constraints, although it is not ideal for their formulation. In contrast, DSPs are comparatively easy to understand, but lack an implementation to validate RDF data. In this paper, we use SPIN as basic validation framework and present a general approach how domain specific constraint languages like DSP can be executed on RDF data using SPARQL as an intermediate language.

Keywords: RDF validation; RDF constraint formulation; RDF constraint validation; Description Set Profiles; DSP; RDF; linked data; semantic web.

1. Introduction

In 2013, the W3C invited experts from industry, government and academia to the RDF Validation Workshop[1] to discuss use cases and requirements for constraint representation and RDF data validation. The following needs are reported:

1. Declarative definition of the structure of a graph for validation and description.
2. Extensible to address specialized use cases.
3. A mechanism to associate descriptions with data.

An important finding is that there are non-functional requirements for data validation in a Linked Data setting, particularly the need to "communicate the constraints against which data is to be validated in a way which is both easy to understand by human beings and discoverable by programs."

Partly as follow-up to the W3C workshop and partly due to further expressed requirements at the Semantic Web in Libraries conference 2013[2], the Dublin Core Metadata Initiative in collaboration with the W3C currently establishes a Task Group for RDF Application Profiles (RDF-AP) that will investigate existing approaches and best-practices, identify possible gaps and propose practical solutions for the representation of application profiles, including the formulation of data constraints[3]. In a heterogeneous environment like the Web, there is not necessarily a one-size-fits-all solution, especially as existing solutions should rather be integrated than replaced, not least to avoid long and fruitless discussions about the "best" approach.

SPARQL and SPIN are powerful and widely used for constraint formulation and validation (Fürber & Hepp, 2010), but constraints formulated as SPARQL queries are not as understandable

[1] RDF Validation Workshop – Practical Assurances for Quality RDF Data. 10-11 September 2013, Cambridge, MA, USA. http://www.w3.org/2012/12/rdf-val/report

[2] SWIB13 – Semantic Web in Libraries, 25 - 27 November 2013, Hamburg, Germany. http://swib.org/swib13/

[3] http://wiki.dublincore.org/index.php/RDF-Application-Profiles

as one wishes them to be. Consider the following example of the simple constraint stating that only dogs are allowed as pets:

```
SELECT ?this ?subope ?object WHERE {
    ?C owl:allValuesFrom :Dog .
     ?C owl:onProperty :hasPet .
     ?C a owl:Restriction .
     ?this rdf:type ?subC . ?subC rdfs:subClassOf* ?C .
     ?this ?subOPE ?object . ?subOPE rdfs:subPropertyOf* :hasPet .
     FILTER NOT EXISTS { ?object rdf:type :Dog . } }
```

This query checks the constraint and returns violating triples, but the actual constraint could be formulated easier using Description Set Profiles[4]:

```
[ a dsp:NonLiteralStatementTemplate;
   dsp:property :hasPet;
   dsp:nonLiteralConstraint [
     dsp:valueClass :Dog;
   ]
]
```

Of course, it can be argued if DSPs are the best possible way to represent constraints. They are, however, familiar to the DCMI community and tailored to the Dublin Core Abstract Model and the Singapore Framework. As stated above, there will probably be more than one constraint language that can be used in an application profile, with DSPs being one of them. This leaves the question, how the validation of data based on different constraint languages can be implemented. Different implementations using different underlying technologies hamper the interoperability of application profiles and a full implementation of several constraint languages is hard to maintain for solution providers. We therefore propose to use SPARQL as intermediate language: constraints in arbitrary languages are transformed to executable SPARQL queries used to validate the data.

This approach obviously requires that all constraint languages can be expressed in SPARQL. We have no formal proof, as use-cases and requirements still are collected and there is neither a complete list of possible constraints nor one of supported constraint languages. However, even if there are constraints that cannot be translated to SPARQL, the subset of supported constraints is certainly large enough to justify the limitation to SPARQL-expressible constraints at least for one class of RDF Application Profiles, comparable to the sublanguages of OWL.

This claim is supported by the fact that SPARQL is already widely used for constraint formulation, as mentioned above. Additionally, Sirin and Tao showed how constraints can be translated to nonrecursive Datalog programs for validation (Sirin & Tao, 2009), while Angles and Gutierrez explained that SPARQL has the same expressive power as nonrecursive Datalog programs (Angles & Gutierrez, 2008).

In this paper, we present our first results regarding the implementation of our approach using SPIN. We will show that besides SPIN, no further dependencies exist. We create a full validation environment based on SPIN that can be used to validate domain specific constraint languages (Section 2). The only limitations are that the constraints have to be expressed in RDF and that the constraint language is expressible in SPARQL. In Section 3, we introduce Description Set Profiles as domain specific constraint language and subsequently describe its implementation in

[4] In RDF-Turtle Syntax, omitting the surrounding description template, for details refer to http:// dublincore.org/documents/dc-dsp

our environment (Section 4). We conclude in Section 5 with a discussion of open questions and an outlook to the next steps.

2. Validation Environment

We use the SPARQL Inferencing Notation (SPIN)[5] to create what we call a validation environment. The overall idea is that we see constraint languages as domain specific languages (hence domain specific constraint languages, DSCL) that are translated and executed on RDF data within our validation environment.

The translation is done once, for instance by the developer of the DSCL, and provided in form of a **SPIN mapping** plus optional **preprocessing instructions**. From a user's perspective, all that is needed is a representation of **constraints using the DSCL** and some **data to be validated** against these constraints. All these resources are purely declarative and provided in RDF or as SPARQL queries. The actual implementation is trivial using SPIN and illustrated in Figure 1.

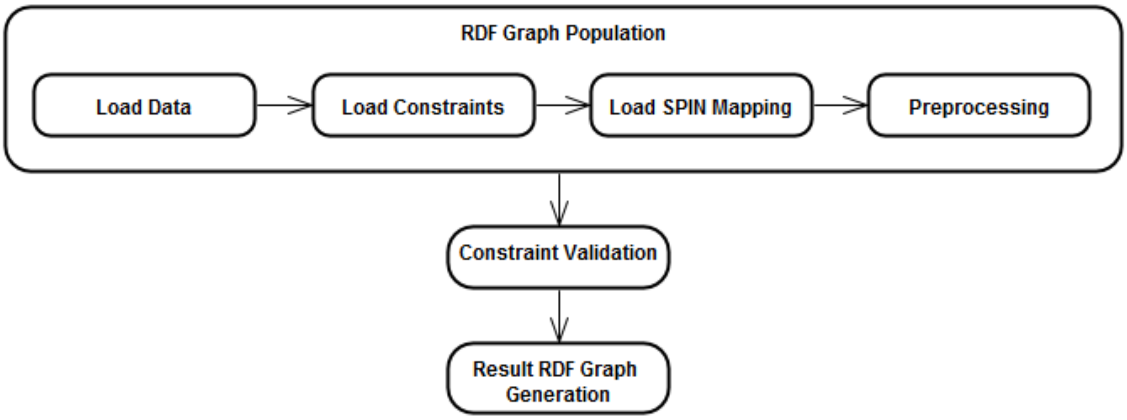


FIG. 1. Constraint Validation Process

First, an RDF graph has to be populated as follows:

1. the **data** is loaded that is to be validated,
2. the **constraints** in the DSCL are loaded,
3. the **SPIN mapping** is loaded that contains the SPARQL representation of the DSCL (see Section 4 for a detailed explanation), and
4. the **preprocessing** is performed, which can for example be provided in the form of CONSTRUCT queries.

When the graph is ready, the SPIN engine checks for each resource in the RDF data if the resource satisfies all defined constraints and generates a result RDF graph containing information about all constraint violations.

With this implementation, there is one obvious limitation of our approach: the DSCL needs an RDF serialization. For DSP, this is the case, but in the future, we would like to support non-RDF languages as well. We will further elaborate on this interesting topic in Section 5.

Connect SPIN to your data. SPIN uses templates for SPARQL queries that are executed on every instance of a given class – for instance `:toValidate`.

Most of the SPIN mapping that has to be created by the DSCL developer consists of such templates that are linked to a class for which the constraints should be evaluated:

[5] http://spinrdf.org/

```
:ToValidate
     spin:constraint
         [ a dsp2spin:StatementTemplates_MinimumOccurrenceConstraint ] .
```

As the mapping is designed to be independent of any actual data, the class `:toValidate` is purely generic. Instead of using such a generic class, it is also possible to link the constraints to `owl:Thing` or `rdfs:Resource`, i.e., to all instances.

Neither of these classes have to be assigned explicitly to instances within the data to be validated. They are either inferred using reasoning or explicitly assigned during the **preprocessing**: a reasonable approach would be to assign `:toValidate` to all classes for which constraints are actually defined – in the case of DSP classes that are linked via `dsp:resourceClass` to a description template; this can be accomplished by a suitable CONSTRUCT query that is executed before the actual validation.

After preprocessing, the data might look like this – with the added generic class in italics:

```
:ArtficialIntelligence
     a swrc:Book, :ToValidate ;
     dcterms:subject :ComputerScience .
```

`:ArtificialIntelligence` denotes a book with the assigned subject "Computer Science."

Mapping from a DSCL to SPARQL. The actual mapping is performed by creating appropriate SPARQL templates for every constraint that is supported in the DSCL, for example a minimum occurrence that is required:

```
dsp2spin:StatementTemplates_MinimumOccurrenceConstraint
     a spin:Template;
     spin:body [
         a sp:Construct ;
         sp:templates (...) ;
         sp:where (...) ] .
```

This is the general structure of a SPIN template representing a SPARQL CONSTRUCT query. We use CONSTRUCT queries to generate descriptions of each constraint violation, for instance:

```
CONSTRUCT {
    _:violation
         a spin:ConstraintViolation ;
         rdfs:label ?violationMessage ;
         spin:violationRoot ?violationRoot ;
         spin:violationPath ?violationPath ;
         spin:violationSource ?violationSource ;
         spin:fix ?violationFix ;
         :severityLevel ?severityLevel }
```

In SPIN, such a CONSTRUCT query is represented in RDF as follows:

```
a sp:Construct ;
sp:templates (
[ sp:subject _:violation ; sp:predicate rdf:type ; sp:object spin:ConstraintViolation ]
[ sp:subject _:violation ; sp:predicate rdfs:label ; sp:object [ sp:varName "violationMessage" ] ]
... ) ;
```

Constraint violation triples (1) provide useful messages explaining the reasons why RDF instances did not satisfy the constraints (`rdfs:label`), (2) contain references to RDF triples causing the constraint violations (`spin:violationRoot`), and (3) include references to the constraints causing constraint violations (`spin:violationSource`). Constraint violation triples give some guidance how to become valid data (`spin:fix`) in order to be able to fix constraint violations. Constraint violations can be classified according to different levels of severity (`:severityLevel`).

These constraint violation triples are generated for each RDF instance which matches against the WHERE clause graph pattern in the SPIN template. The SPARQL variable this represents the current RDF resource for which the constraint is checked.

As the mapping of a DSCL is independent of a concrete constraint specification, all constraints are generally linked to all instances (of the generic class, if applicable). Therefore, the WHERE clause of the template always have to restrict on a class for which the constraint was actually defined, for example in the case of DSP via the resource class:

```
WHERE { ?this rdf:type ?resourceClass . }
```

As for the CONSTRUCT part of the query, SPIN represents the WHERE clause in RDF as well:

```
[ sp:subject [ sp:varName "this" ] ;
     sp:predicate rdf:type ; sp:object [ sp:varName "resourceClass" ] ]
```

With this framework, we have all we need to implement our own DSCL, Description Set Profiles, which we will briefly introduce in the next section. Full examples for SPIN mappings are provided afterwards in Section 4.

3. DSP as Domain Specific Constraint Language

The Singapore Framework[6] is a framework for designing metadata and for defining Dublin Core Application Profiles (DCAP). The framework comprises descriptive components that are necessary or useful for documenting DCAPs. A DCAP is a means to assemble and to customize components from different independently created metadata standards within the context of a specific community, application, and domain[7].

The DCMI Abstract Model (DCAM)[8] with its Description Set Model (DSM) forms the basis of Dublin Core metadata. While the DSM is highly related to RDF, it differs in some aspects worth mentioning. Table 1 shows the mappings from DSM elements to RDF triples, according to DC-RDF, the recommendation how Dublin Core metadata is represented in RDF[9].

TABLE 1: DSM in RDF

DSM	RDF
Description Set	RDF graph (containing description RDF graphs)
Description	RDF graph
Resource	RDF subject: DSM resource URI (or blank node) (root of description RDF graph)

[6] http://dublincore.org/documents/singapore-framework/

[7] cf. http://dublincore.org/documents/profile-guidelines/

[8] http://dublincore.org/documents/2007/06/04/abstract-model/

[9] http://dublincore.org/documents/dc-rdf/

Statement	RDF subject: DSM resource RDF predicate: RDF property RDF object: DSM value (surrogate)
Non-Literal Value Surrogate	DSM value URI (or blank node)
Vocabulary Encoding Scheme	RDF subject: DSM value RDF predicate: `dcam:memberOf` RDF object: DSM vocabulary encoding scheme
Value String	RDF subject: DSM value RDF predicate: `rdf:value` RDF object: RDF Literal (DSM value string) (RDF plain literal or RDF typed literal)
Literal Value Surrogate	DSM value is RDF literal (RDF plain literal or RDF typed literal)
Value String Language	Language tag of RDF literal
Syntax Encoding Scheme	RDF datatype of RDF typed literal

A Description Set Profile (DSP)[10] contains constraints on the data within a DCAP, i.e., a DSP restricts valid descriptions of resources in a description set. Consider the following example of a DSP:

```
:bookDescriptionTemplate
    a dsp:DescriptionTemplate ;
    dsp:standalone "true"^^xsd:boolean ;
    dsp:minOccur "1"^^xsd:nonNegativeInteger ; dsp:maxOccur "infinity"^^xsd:nonNegativeInteger ;
    dsp:resourceClass swrc:Book ;
    dsp:statementTemplate [
        a dsp:NonLiteralStatementTemplate ;
        dsp:minOccur "1"^^xsd:nonNegativeInteger ; dsp:maxOccur "5"^^xsd:nonNegativeInteger ;
        dsp:property dcterms:subject ;
        dsp:nonLiteralConstraint [
            a dsp:NonLiteralConstraint ;
            dsp:descriptionTemplate :subjectDescriptionTemplate ;
            dsp:valueClass skos:Concept ;
            dsp:valueURIOccurrence "mandatory"^^dsp:occurrence ;
            dsp:valueURI :ComputerScience, :SocialScience, :Librarianship ;
            dsp:vocabularyEncodingSchemeOccurrence "mandatory"^^dsp:occurrence ;
            dsp:vocabularyEncodingScheme :BookSubjects ;
            dsp:valueStringConstraint [
                a dsp:ValueStringConstraint ;
                dsp:minOccur "1"^^xsd:nonNegativeInteger ; dsp:maxOccur "1"^^xsd:nonNegativeInteger ;
                dsp:literal "Computer Science"@en , "Computer Science"^^xsd:string ;
                dsp:literal "Social Science"@en , "Social Science"^^xsd:string ;
                dsp:literal "Librarianship"en , "Librarianship"^^xsd:string ;
                dsp:languageOccurrence "optional"^^dsp:occurrence ;
                dsp:language "en"^^xsd:language ;
                dsp:syntaxEncodingSchemeOccurrence "optional"^^dsp:occurrence ;
                dsp:syntaxEncodingScheme xsd:string ] ] ] .
```

A DSP consists of `dsp:DescriptionTemplates` that put constraints on instances of a certain class, denoted by `dsp:resourceClass`. The constraints can either be constraints on the description itself, e.g., a minimum occurrence of instances of this class. Additionally, constraints on single properties can be defined within a `dsp:StatementTemplate`. The example above contains all but one of the 23 constraints defined in DSP (except the sub-property constraint; the 5 literal value constraints can be used for value strings as well).

[10] http://dublincore.org/documents/2008/03/31/dc-dsp/

The DSM description template `:bookDescriptionTemplate` describes DSM resources of the type `swrc:Book` (referenced by `dsp:recourceClass`). `swrc:Book` resources are allowed to occur standalone (`dsp:standalone`), i.e. without being the value of a property. Books must occur at least once (`dsp:minOccur`) and may appear multiple times (`dsp:maxOccur`) in the DSM description set (the RDF graph). The `dsp:NonLiteralStatementTemplate` restricts books to have at least 1 (`dsp:minOccur`) and at most 5 (`dsp:maxOccur`) `dcterms:subject` (`dsp:property`) relationships to DSM non-literal value surrogates which are further described by the `dsp:NonLiteralConstraint`.

The DSM values have to be of the class `skos:Concept` (`dsp:ValueClass`) and are further described in a dedicated DSM description template (referenced by `dsp:descriptionTemplate`). A value URI must be given (`dsp:valueURIOccurrence`) for DSM values and allowed value URIs (`dsp:valueURI`) are `:ComputerScience`, `:SocialScience`, and `:Librarianship`. Controlled vocabularies (like `:BookSubjects`) are represented as `skos:ConceptSchemes` in RDF and as `dsp:VocabularyEncodingSchemes` in DSM. If DSM vocabulary encoding schemes must be stated (`dsp:vocabularyEncodingSchemeOccurrence`), they have to contain the DSM values. In this case, DSM values are classified as `skos:Concepts` and are related to `skos:ConceptSchemes` via the object properties `skos:inScheme` and `dcam:memberOf` (see RDF data above).

The DSM values must be represented as exactly one (`dsp:minOccur` and `dsp:maxOccur` - line 20) of the given three DSM value strings (`dsp:literal`). The language tag en (`dsp:language`) as well as the RDF datatype xsd:string (`dsp:syntaxEncodingScheme`) may be stated (`dsp:languageOccurrence` and `dsp:syntaxEncodingSchemeOccurrence`) for DSM value strings.

An example for RDF data satisfying all these constraints for resources of the type `swrc:Book` would be:

```
:ArtficialIntelligence
    a swrc:Book , :ToValidate ;
    dcterms:subject :ComputerScience .
:ComputerScience
    skos:Concept , :ToValidate ;
    dcam:memberOf :BookSubjects ;
    skos:inScheme :BookSubjects ;
    rdf:value "Computer Science"@en .
:BookSubjects
    a skos:ConceptScheme , :ToValidate .
```

4. Mapping of DSP Constraints to SPIN

After the introduction of the general approach in Section 2, we now present a concrete example of a SPIN mapping for a DSP constraint: the DSP statement template constraint 'Minimum Occurrence Constraint' (6.1) restricts the minimum number of times the given statement must appear in the enclosing description.

This constraint is implemented by the following SPARQL query which is then represented in SPIN RDF and linked to our generic class *`:ToValidate`*:

```
CONSTRUCT {
    _:violation
        a spin:ConstraintViolation ;
        rdfs:label ?violationMessage ;
        spin:violationRoot ?this ;
        spin:violationSource dsp:minOccur }
WHERE {
    ?this rdf:type ?resourceClass .
```

```
     ?descriptionTemplate rdf:type dsp:DescriptionTemplate .
     ?descriptionTemplate dsp:resourceClass ?resourceClass .
     ?descriptionTemplate dsp:statementTemplate ?statementTemplate .
     ?statementTemplate dsp:minOccur ?minOccurStatement .
     ?statementTemplate dsp:property ?property .
     BIND ( ( spl:objectCount ( ?this, ?property ) ) AS ?cardinalityStatement ) .
     FILTER ( cardinalityStatement < ?minOccurStatement ) .
     BIND ( (
         fn:concat('cardinality of ', ?property, ' ( ', ?cardinalityStatement, ' )
         < mininum cardinality of ', ?property, ' ( ', ?minOccurStatement, ' )' ) )
         AS ?violationMessage ) . }
```

It can be seen that the WHERE clause is used to "detect" constraint violations. First, a graph is matched that contains the instance data (using `?this` as instance variable) and the applicable constraint formulation from the DSP (linked to the instance via `dsp:resourceClass`). The cardinality of the property in question is added. The actual validation is implemented by the FILTER that identifies only instances that violate the constraint.

In this example, we create a violation message (`?violationMessage`) that can be displayed to the user, together with the URI of the instance (?this as `spin:violationRoot`) and the violated constraint (`dsp:minOccur` as `spin:violationSource`).

According to our DSP, if a resource in the RDF data

1. is assigned to the class `swrc:Book` (line 5), and
2. has no `dcterms:subject` relationships (line 8 and 9),

then the following constraint violation triple is generated:

```
_:violation
     a spin:ConstraintViolation ;
     rdfs:label
        'cardinality of dcterms:subject ( 0 ) < mininum cardinality of dcterms:subject ( 1 )' ;
     spin:violationRoot :IntroductionToAlgorithms ;
     spin:violationSource dsp:minOccur .
```

This example demonstrates how a DSP constraint is implemented in SPARQL. In the same manner, most other constraints can be implemented as well, although often the mapping gets substantially longer and more complex. There are, however, constraints that cannot be implemented at all, in the case of DSP for example the literal value constraint *Syntax Encoding Scheme Constraint (6.5.4)*. It determines whether DSP syntax encoding schemes (RDF datatypes) are allowed for RDF literals, which can be 'mandatory', 'optional', or 'disallowed'.

This type of constraint cannot be validated as RDF literals always have associated datatype IRIs. If there is no datatype IRI and no language tag explicitly stated, the datatype of an RDF literal is implicitly `xsd:string`. If there is a language tag, the datatype is implicitly `rdf:langString`. Fortunately this constraint can be replaced by an equivalent constraint using *Syntax Encoding Scheme List Constraint (6.5.5)* which restricts the allowed DSP syntax encoding schemes and which is fully implemented in the SPIN mapping for DSP.

5. Conclusion and Future Work

With our approach, we were able to fully implement Description Set Profiles, apart from the exception noted above. The implementation can be tested at `http://purl.org/net/rdfval-demo`. In this paper, we describe our general approach and demonstrated its applicability to Description Set Profiles. In particular, we use SPIN as basis to define a validation environment in which domain specific constraint languages – like DSP – can be implemented by representing them in SPARQL. The approach is particularly appealing as it has only one dependency being SPIN. The implementation of the DSCL is fully declarative, consisting of a SPIN mapping in RDF and preprocessing instructions in form of SPARQL CONSTRUCT queries – which can also be represented in RDF using SPIN. It is therefore possible to link the applicable constraints in a given DSCL to an application profile, as well as the SPIN mapping and the preprocessing instructions. All that is needed to validate data according to this application profile without the need for a DSCL-specific validator. Our approach therefore fulfills an important requirement for RDF Application Profiles.

A limitation of our approach are constraints that cannot be expressed in SPARQL, as for example the Syntax Encoding Scheme Constraint of DSP. In this case this is an artefact resulting from the way how RDF is implemented. There are most certainly other cases, but we argue that our approach is nevertheless useful for the majority of constraints in the majority of DSCLs. We propose, however, to document such missing constraints clearly as part of the DSCL so that users can deal with it.

Our approach is currently limited to DSCLs that are expressible in RDF. This is not necessarily a problem – the data and the data models are in RDF, so at least it is consistent – but it might be sub-optimal regarding readability and understandability of the constraints and for now excludes many existing DSCLs. We therefore plan to investigate this issue further as part of our future work. Another interesting topic is the testing of the SPIN mappings, for which test data together with expected outcomes could be provided in a certain form. Our next steps include the application to further constraint languages, first and foremost OWL2, which is already used by many to formulate constraints. The DSP mapping is developed and maintained at `https://github.com/dcmi/DSP-SPIN-Mapping`.

Acknowledgements

Kai Eckert is funded by the European Commission within the DM2E project (`http://dm2e.eu`)

References

Angles, R., & Gutierrez, C. (2008). The expressive power of SPARQL. In Proceedings of the 7th International Semantic Web Conference (ISWC2008) (pp. 114–129).

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

Sirin, E., & Tao, J. (2009). Towards integrity constraints. In Proceedings of the Workshop on OWL: Experiences and Directions, OWLED 2009.

Describing Theses and Dissertations Using Schema.org

Jeff Mixter
OCLC, USA
mixterj@oclc.org

Patrick OBrien
Montana State University, USA
patrick.obrien4@montana.edu

Kenning Arlitsch
Montana State University, USA
kenning.arlitsch@montana.edu

Abstract

This report discusses the development of an extension vocabulary for describing theses and dissertations, using Schema.org as a foundation. Instance data from the Montana State University ScholarWorks institutional repository was used to help drive and test the creation of the extension vocabulary. Once the vocabulary was developed, we used it to convert the entire ScholarWorks data sample into RDF. We then serialized a set of three RDF descriptions as RDFa and posted them online to gather statistics from Google Webmaster Tools. The study successfully demonstrated how a data model consisting of primarily Schema.org terms and supplemented with a list of granular/domain specific terms can be used to describe theses and dissertations in detail

Keywords: Schema.org; RDF; linked data; institutional repositories; semantic web; search engine optimization; data modeling.

1. Introduction

As academic institutions realize the value of their intellectual output, well-organized and discoverable institutional repositories are increasingly viewed as strategic assets. The intellectual output of an academic institution is diverse and ranges from student theses and dissertations to conference proceedings, presentations, books, journal articles, and the datasets that support research conclusions. It is crucial for purposes of discovery to publish the metadata in a format that is easily understood, consumed and indexed by search engines and other machine-based data aggregators.

This project builds on research whose initial aim was to improve visibility of digitized special collections in commercial search engines, and was partially funded by the Institute of Museum and Library Services (IMLS). The first phases of research were successful in developing search engine optimization (SEO) strategies and methods, and led to the publication of a book (Arlitsch & OBrien, 2013). Beyond digitized special collections the research also revealed that institutional repositories (IRs) pose unique and complex problems to scholarly search engines, and as a result many IRs were not being consistently harvested and indexed. The project described in this report examines a specific subset of IR content, theses and dissertations. The scope of the project was to create a set of extension terms for Schema.org[1] that can be used to describe theses and dissertations and to create a process model that explains how we converted the existing Montana State University Dublin Core metadata into Linked Data. Following this proof of concept, we plan to explore how to integrate the new vocabulary into existing IRs so that they can provide search engines with more meaning and context, ultimately resulting in more accurate search results for users.

1.1. Data Sample

We used the Montana State University ScholarWorks IR dataset to drive and validate the modeling process that expanded and implemented the Schema.org vocabulary. This approach provided the group with a multitude of rich modeling examples and use cases but it also helped

[1] http://schema.org

keep the process of modeling firmly grounded in the requirements presented by the data. The ScholarWorks dataset that was used for the study was a collection of student theses and dissertations. There were 1909 records in the sample, which had originally been described using Dublin Core (DC) and, where necessary, additional DC extensions for granular details. It should be noted that prior to use in this study, the ScholarWorks metadata was cleaned up to ensure that all of the fields were populated with information, where appropriate, and that the fields were used according to their proper definitions. This prior work mitigated the need to perform an initial review and cleanup in order to use the data, but IR managers who plan to implement structured metadata should be aware that this cleanup is a crucial first step.

2. Extension Vocabulary Development

In our initial review of the dataset, we tried to use existing vocabularies to describe theses and dissertations. It became evident when reviewing the sample data extracted from ScholarWorks that existing vocabularies alone were not robust enough to fully describe the items. Application Profiles were an attempt by the larger metadata community to develop a set of vocabulary terms that can be used within a specific context to describe unique items. The idea was that a metadata schema could be developed from a variety of existing schemas, modified if needed and then used to describe a unique set of items within the context of a specific application or domain (Heery & Patel, 2000). Sir Tim Berners-Lee referred to this same type of modeling as "cherry-picking" at the Gov 2.0 Expo in 2010, suggesting that nearly all of the vocabulary terms that one would need to describe an item already exist (Berners-Lee, 2010). The work around application profiles was recently restarted within the context of developing RDF application profiles. A DCMI Task Group has begun to investigate how RDF application profiles could be created and used to help with data validation.[2] An early example of picking and choosing RDF terms from a variety of vocabularies can be found in the British Data Model (Hodson, Deliot, Danskin, Rosie & Ashton, 2012). In this model, terms are taken from fifteen different vocabularies and combined to form a comprehensive model for describing bibliographic items.

We used the same approach to develop the extension vocabulary for the theses and dissertations sample set. Below is a table showing the vocabularies that we used.

TABLE 1: Vocabularies used in the project

Vocabularies used in the project	
Prefix	Namespace
schema	http://schema.org
dcterms	http://purl.org/dc/terms/
pto	http://www.productontology.org/id/
rdfs	http://www.w3.org/2000/01/rdf-schema#
mont	http://purl.org/montana-state/library/

In addition to Table 1, we created and published a VoID dataset description.[3] It includes information about the sample datasets, including dataset statistics. The extension vocabulary that we developed was not designed to be prescriptive. Rather, it was meant to be used with the entire Schema.org vocabulary. In this sense, our extension vocabulary provides a descriptive way for rationalizing existing descriptions of theses and dissertations as Linked Data without adding any constraints or validation requirements. As Linked Data graphs continue to grow in size, validation will obviously become an important topic and requirement for systems/services. Over the next few years, it will be interesting to observe the path that the RDF Application Profile Task Group

[2] http://wiki.dublincore.org/index.php/RDF_Application_Profiles

[3] http://purl.org/montana-state/scholarworks/sampledataset

takes in dealing with validation requirements. The full list of extension classes and properties are available online.[4]

2.1. Classes

The new classes we developed for the extension vocabulary were divided into two unique categories. The first category included class extensions that were used to add a more granular description of the item being described. The labels for these classes were derived from the 'Appendix III – Types' controlled vocabularies used in the Citation Style Language.[5] Table 2 lists the first category of classes.

TABLE 2: Citation Style Language terms

Extension Classes derived from Citation Style Language terms
mont:JournalArticle
mont:MagazineArticle
mont:NewspaperArticle
mont:Bill
mont:Chapter
mont:ConferencePaper
mont:Entry
mont:Figure
mont:Graphic
mont:Interview
mont:LegalCase
mont:Legislation
mont:Manuscript
mont:MusicalScore
mont:Pamphlet
mont:Patent
mont:PersonalCommunication
mont:Report
mont:Speech
mont:Thesis
mont:Treaty

The second category of classes that was developed for the extension vocabulary included terms that were not covered by existing popular vocabularies but were required for the description of theses and dissertations. Table 3 lists the second category of classes that were created for the extension vocabulary.

TABLE 3: Extension Classes not covered by other vocabularies

Extension Class
mont:AcademicDepartment
mont:Collection
mont: School
mont:Concept
mont:DigitalCollection
mont:DoctoralThesis
mont:EtdCommittee
mont:InstitutionalRepository
mont:MasterThesis
mont:ScholarlyWork
mont:SpecialCollection

[4] http://purl.org/montana-state/library

[5] http://citationstyles.org/downloads/specification.html#appendix-iii-types

A diagram of the classes and relationships used in the project can be found in Appendix I.

2.2. Properties

Although Schema.org has a wide variety of properties, the ScholarWorks instance data helped us identify use cases that required more granular terms to properly describe the item. We were able to create relationships between entities that were otherwise mashed together in the Dublin Core records. Figure 1 illustrates how we were able to identify individual people and committees and also define how they were related to each other.

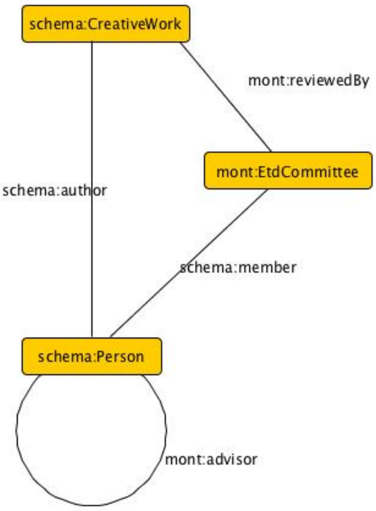


FIG 1: Relationships derived from DC records

Table 4 contains all of the properties that were created for the extension vocabulary as well as the type of Web Ontology Language (OWL) property that should be interpreted for each.

TABLE 4: List of Properties and OWL equivalencies

Extension vocabulary property	Object or Data property
mont:associatedDepartment	Object
mont:associatedSchool	Object
mont:adviser	Object
mont:campus	Object
mont:committeeChair	Object
mont:committeeMember	Object
mont:curates	Object
mont:facultyMember	Object
mont:hadDepartment	Object
mont:hasEtdCommittee	Object
mont:hasLibrary	Object
mont:reviewedBy	Object
mont:callNumber	Data
mont:degreeGrantedForCompletion	Data
mont:degreeGranted	Data
mont:firstPage	Data
mont:lastPage	Data

3. Testing And Implementing The Model

After the model was developed, the entire ScholarWorks dataset was converted into Linked Data using a modified version of OpenRefine[6] called LODRefine.[7] Once the data were imported into LODRefine, a variety of data cleanup tasks were conducted and finally the Schema.org and extension vocabulary were imported and used to generate Linked Data. The first major cleanup task was to separate cells that contained multiple values into individual cells. After completing the cleanup we attempted to reconcile named entities to existing Linked Data datasets. We queried several datasets, including LCSH, VIAF and DBpedia. The most successful matching came from values that were included in the 'subjects', 'subjects.lcsh' and 'coverage.spatial' fields. The 'subject.lcsh' terms had a particularly high match rate (78% match to LCSH URIs) while the other fields matched at a lower rate (40% matched to DBpedia.org). The one problem with querying LCSH terms was that there were many pre-coordinated headings. Since the LCSH Linked Data dataset only includes terms that are part of the LCSH Authority files, there were quite a few terms that did not match up correctly. A solution to this problem would be to coin local URIs for the pre-coordinated headings and then include dc:hasPart or rdfs:seeAlso properties pointing out to the individual LCSH URIs that are referenced in the compound heading.

For the named entities that did not reconcile to the aforementioned datasets, local URIs were coined. These URIs followed a set pattern and then used the string value of the field as the identifier token. Figure 2 is an example of one of the URIs that was created when we could not match it to an existing Linked Data dataset.

http://scholarworks.montana.edu/doc/entities.html#person/Angie_Keesee

FIG 2: Sample URI coined for string value

More information about how to clean up dirty data and generate Linked Data using OpenRefine can be found in (Vorborgh & De Wild, 2013). In order to publish the Linked Data in a web-friendly serialization and to begin to test how much structured data search engines can mine, we converted three of the descriptions into RDFa and published them on ScholarWorks.[8] For all of the entities that did not have existing metadata records, such as people, places, organizations, etc, a single HTML page was generated that has a list of entity descriptions. The page is anchored with the URI tokens that appear after the #, so if one of these 'extra entity' URIs is resolved in the browser it will position the user in the appropriate portion of the page. The list can be found at Montana Scholar Works.[9]

3.1. Instance Data Example

In order to give a better understanding of the results of the modeling, this section walks through one of the sample records that was converted into Linked Data. The full RDFa description of this record is available online.[10] Figure 3 on the following page provides a graphic representation of the terms used to describe the item. The sample pictured in Figure 3 is also expressed in Turtle in Appendix II. The diagram does not list all of the properties and classes that can/should be used to describe theses and dissertations. A complete list of all of the terms used in the sample collection can be found in the Appendix III.

[6] http://openrefine.org/

[7] http://code.zemanta.com/sparkica/

[8] http://scholarworks.montana.edu/doc/index.html).

[9] http://scholarworks.montana.edu/doc/entities.html

[10] http://scholarworks.montana.edu/doc/SampleWork1.html

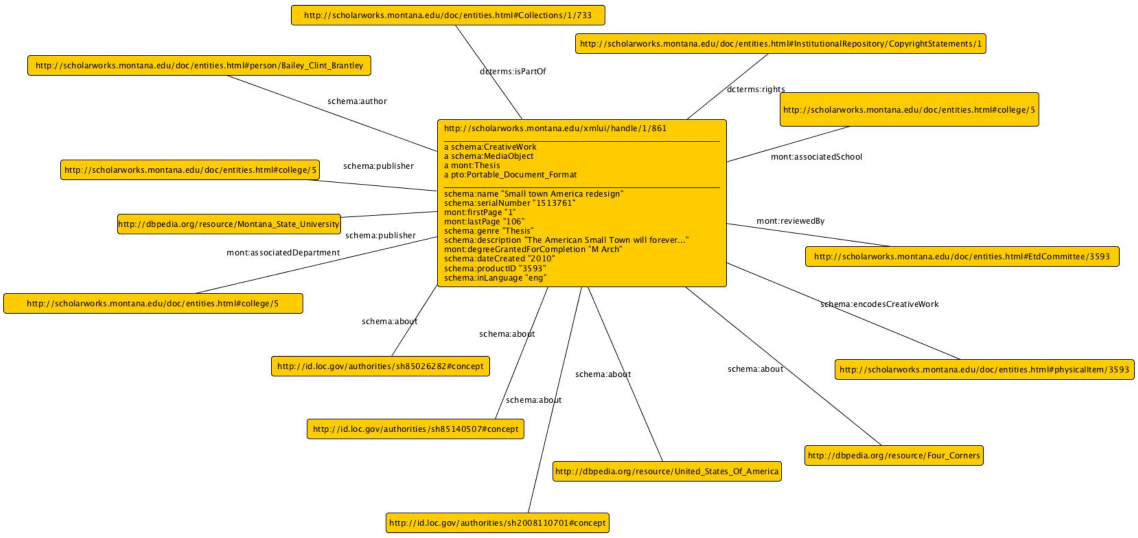

http://scholarworks.montana.edu/doc/entities.html#Collections/1/733
http://scholarworks.montana.edu/doc/entities.html#InstitutionalRepository/CopyrightStatements/1
http://scholarworks.montana.edu/doc/entities.html#person/Bailey_Clint_Brantley
dcterms:isPartOf
dcterms:rights
schema:author
http://scholarworks.montana.edu/doc/entities.html#college/5
http://scholarworks.montana.edu/xmlui/handle/1/861
a schema:CreativeWork
a schema:MediaObject
a mont:Thesis
a pto:Portable_Document_Format
schema:name "Small town America redesign"
schema:serialNumber "1513761"
mont:firstPage "1"
mont:lastPage "106"
schema:genre "Thesis"
schema:description "The American Small Town will forever..."
mont:degreeGrantedForCompletion "M Arch"
schema:dateCreated "2010"
schema:productID "3593"
schema:inLanguage "eng"
mont:associatedSchool
http://scholarworks.montana.edu/doc/entities.html#college/5
schema:publisher
http://dbpedia.org/resource/Montana_State_University
schema:publisher
mont:reviewedBy
http://scholarworks.montana.edu/doc/entities.html#EtdCommittee/3593
mont:associatedDepartment
http://scholarworks.montana.edu/doc/entities.html#college/5
schema:encodesCreativeWork
schema:about
schema:about
http://id.loc.gov/authorities/sh85026282#concept
schema:about
schema:about
http://scholarworks.montana.edu/doc/entities.html#physicalItem/3593
schema:about
http://id.loc.gov/authorities/sh85140507#concept
http://dbpedia.org/resource/Four_Corners
http://dbpedia.org/resource/United_States_Of_America
http://id.loc.gov/authorities/sh2008110701#concept

FIG 3: Graphical representation of sample item

4. Conclusion

We were able to successfully map the theses and dissertations metadata into Schema.org and, when needed, supplemented existing Dublin Core fields with terms we created as part of an extension vocabulary for Schema.org. The extension terms followed the same standards and practices as those in Schema.org and every attempt was made to position extension terms as sub-classes or sub-properties of existing Schema.org terms. The project has thus far successfully developed an extension vocabulary to describe theses and dissertations and show how to apply the vocabulary to existing metadata. Since modeling is an iterative process, the next step in the project will be to apply the vocabulary to more sets of theses and dissertations and make additions/changes. We also plan to publish more RDFa and begin to track the amount of structured data that is harvested by search engines using tools such as Google Webmaster Tools.[11]

Acknowledgements

This work could not have been completed without the help and assistance of Dr. Jean Godby and Jeff Young of OCLC Research.

References

Arlitsch, K., & OBrien, P. S. (2013). Improving the visibility and use of digital repositories through SEO. Chicago: ALA TechSource, an imprint of the American Library Association. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=578551

Berners-Lee, T. (2010). Open, Linked Data for a Global Community. Washington D.C. Retrieved from https://www.youtube.com/watch?v=ga1aSJXCFe0&feature=player_embedded

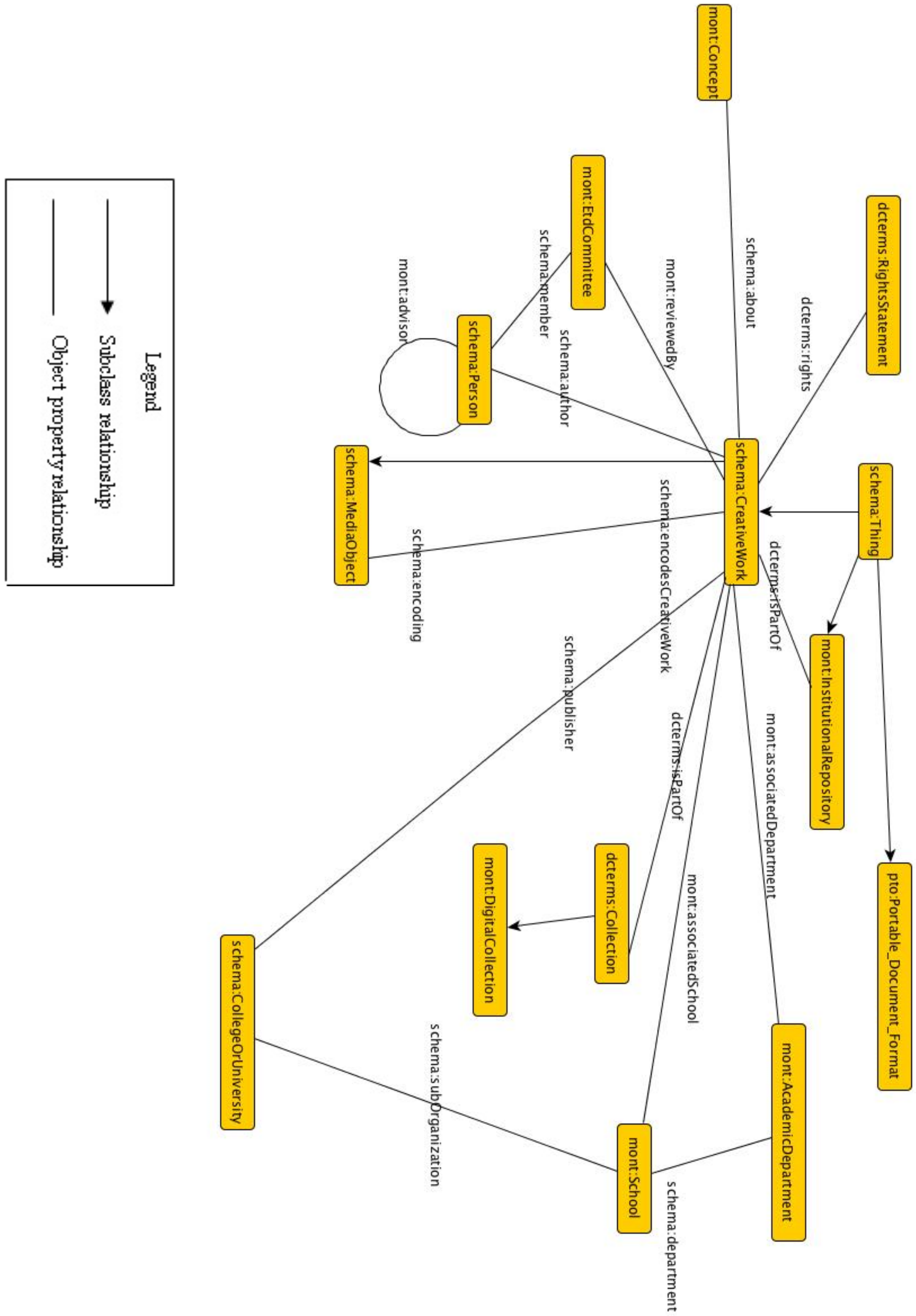
Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. Ariadne, 25, 27-31.

Hodson, T., Deliot, C., Danskin, A., Rosie, H., & Ashton, J. (2012). British Data Model – Book. Retrieved from http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf

Vorborgh, R., & De Wilde, M. (2013). Using OpenRefine: The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web. Packt Publishing Ltd.

[11] https://www.google.com/webmasters/tools/home?hl=en

Appendix I: Visual graph of the vocabulary terms used



Appendix II: Sample data serialized as Turtle

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ns1: <http://purl.org/montana-state/library/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xhv: <http://www.w3.org/1999/xhtml/vocab#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://scholarworks.montana.edu/xmlui/handle/1/861> a schema:CreativeWork,
        schema:MediaObject, mont:Thesis,
        <http://www.productontology.org/id/Portable_Document_Format> ;
    dcterms:isPartOf <http://scholarworks.montana.edu/doc/entities.html#Collections/1/733> ;
    dcterms:rights <http://scholarworks.montana.edu/doc/entities.html#InstitutionalRepository/CopyrightStatements/1> ;
    ns1:associatedDepartment <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
    ns1:associatedSchool <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
    ns1:degreeGrantedForCompletion "M Arch" ;
    ns1:firstPage "1" ;
    ns1:lastPage "106" ;
    ns1:reviewedBy <http://scholarworks.montana.edu/doc/entities.html#EtdCommittee/3593> ;
    schema:about <http://dbpedia.org/resource/Four_Corners>,
        <http://dbpedia.org/resource/United_States_Of_America>,
        <http://id.loc.gov/authorities/sh2008110701#concept>,
        <http://id.loc.gov/authorities/sh85026282#concept>,
        <http://id.loc.gov/authorities/sh85140507#concept> ;
    schema:author <http://scholarworks.montana.edu/doc/entities.html#person/Bailey_Clint_Brantley> ;
    schema:dateCreated "2010" ;
    schema:description "The American Small Town will forever have a place in the undertones of American culture and
in the American psychy. The small town has become an identifing piece of the fabric that the overall American Society
as a whole uses to project its own image, not only to the world but to its self. This study is an examination of key
elements of the American Small town and an exploration into why these places are disappearing. The study goes on to
utilize this information to derive a plan for a small town that is free of modern day plights, such as sprawl and
redundency. In the end, it proposes a plan for the community of Four Corners, M.T. This case study re-design is an
example of how small communities can be shaped early on to prevent waste, maximize efficiency and quality of life." ;
    schema:encodesCreativeWork <http://scholarworks.montana.edu/doc/entities.html#physicalItem/3593> ;
    schema:genre "Thesis" ;
    schema:inLanguage "eng" ;
    schema:name "Small town America [electronic resource] : a re-design / by Clint Brantley Bailey.",
        "Small town America redesign" ;
    schema:productID "3593" ;
    schema:publisher <http://dbpedia.org/resource/Montana_State_University>,
        <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
    schema:serialNumber "1513761" .
```

Appendix III: List of classes and properties used in the study

Classes
schema:Intangible
schema:Person
schema:Organization
schema:CreativeWork
schema:CollegeOrUniversity
schema:EducationalOrganization
schema:MediaObject
pto:Portable_Document_Format
dcterms:RightsStatement
dcterms:Collection
mont:Concept
mont:EtdCommittee
mont:School
mont:InstitutionalRepository
mont:DigitalCollection
mont:AcademicDepartment
Object Properties
schema:subOrganization
schema:encoding
schema:author
schema:member
schema:encodesCreativeWork
schema:about
schema:department
schema:publisher
dcterms:isPartOf
dcterms:rights
mont:advisor
mont:associatedDepartment
mont:associatedSchool
mont:reviewedBy
Data Properties
schema:genre
schema:dateCreated
schema:inLanguage
schema:url
schema:serialNumber
schema:name
schema:productID
schema:description
mont:firstPage
mont:lastPage
mont:degreeGrantedForCompletion
mont:degreeGranted
rdfs:label

Metadata Praxis

Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata

Chunqiu Li
Graduate School of Library, Information and Media Studies,
University of Tsukuba, Japan
licq.chunqiu@gmail.com

Shigeo Sugimoto
Faculty of Library, Information and Media Science,
University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp

Abstract

Provenance description is necessary for long-term preservation of digital resources. Open Archival Information System (OAIS) and Preservation Metadata: Implementation Strategies (PREMIS), which are well-known standards designed for digital preservation, define descriptive elements for digital preservation. Metadata has to be preserved as well as primary resource in order to keep the primary resources alive. However, due to the changing technology and information context, not only primary digital resources but also metadata are at risk of damage or even loss. Thus, metadata preservation is important as well as preservation of primary digital resources. Metadata preservation is a rather new research topic but critical for keeping metadata about preserved resources consistently over time. This paper focuses on provenance as an important issue in digital preservation. It discusses provenance description based on two major metadata standards—PROV and PREMIS. The goal of this study is to clarify a model for describing provenance for metadata preservation. This paper first describes some well-known standards—OAIS, PREMIS, PROV, and so forth, and then discusses a novel model of provenance description based on the PROV Ontology (PROV-O) and PREMIS OWL Ontology. The paper gives provenance description examples using PROV-O and PREMIS OWL Ontology respectively. Based on analysis and mapping among the basic classes of the PROV-O and PREMIS OWL Ontology, we propose an integrated, merged model. We discuss metadata schema provenance and some other open issues.

Keywords: digital provenance; metadata provenance; metadata longevity; PROV; PREMIS

1. Introduction

Metadata plays crucial roles in long-term use of digital resources and digital preservation. Damage or loss of metadata over time may cause serious problems in the long-term use of digital resources. Metadata schema changes may cause inconsistency in the use of metadata, which is also a risk for the long-term use of digital resources. Due to the high cost of re-creation of metadata, longevity of metadata is an important issue for long-term use of digital resources. Metadata schema, which defines a set of terms, structure of metadata instances and some related characteristics of metadata instances, has to be maintained as well as the metadata instances over time.

Provenance information is necessary for long-term use and preservation of digital resources. Provenance is a fundamental principle of archives (Pearce-Moses, 2005) and keeping provenance of every archived item is a fundamental archival function. Open Archival Information System (OAIS) and Preservation Metadata: Implementation Strategies (PREMIS) are widely accepted standards for digital preservation. They include provenance descriptions as primary information. Both OAIS and PREMIS state the importance of provenance description for preservation (Consultative Committee for Space Data System, 2012; PREMIS Editorial Committee, 2012).

As provenance is a general concept, provenance description is not limited to preservation of digital objects. There are several standards for provenance description such as PROV developed by the World Wide Web Consortium (W3C). PROV is defined as a general, high-level standard for provenance, whereas provenance descriptions in OAIS and PREMIS are defined for preservation of information resources. The primary goal of this paper is to study a model for describing provenance of metadata by combining PROV and PREMIS.

This study is primarily aimed at understanding the underlying model for the provenance of metadata for long-term use of metadata—in other words, the interoperability of metadata over time. Metadata preservation is purposed to assure the persistent availability, understandability, and usability of metadata. To make metadata interpretable correctly in the future context is a main goal of metadata preservation. Longevity of digital objects is well known as a crucial issue for the further progress of the networked information society. The technology standards for longevity of digital objects are applicable to the metadata instances because the metadata instances are mostly, but not necessarily, digital objects—e.g., an XML text file and an Excel file. Longevity of digital objects does mean that the objects can be correctly rendered over time. However, it does not necessarily mean that future users can properly understand the content of the object. For example, a table stored in an Excel file may be rendered over time but the attributes of the table cannot be properly understood without proper description of the meaning of the attributes and values. This table example shows a typical problem in metadata preservation—metadata as a digital object may be preserved; but metadata as a semantically meaningful entity may be lost. Even if a metadata instance is encoded in XML and stored in a plain-text file, semantics of XML elements may be lost if the meanings of the tags in the XML text are not properly preserved. Thus, preservation of metadata is not same as preservation of digital objects.

Metadata registries, which store the definitions of metadata terms and controlled vocabularies and provide them over the Internet, have crucial roles in making the metadata terms and controlled vocabularies usable across communities and over time. Moreover, maintaining application profiles is a crucial function for long-term use of metadata. However, management and use of provenance information of the metadata terms and vocabularies has not been discussed except for versioning and its control. Provenance of application profiles has been neither well discussed nor well recognized.

Based on this understanding about state-of-the-art of metadata provenance, this paper discusses a basic model for metadata provenance. The proposed model is defined based on PROV Ontology (PROV-O) and PREMIS OWL ontology. The rest of this paper is organized as follows. Section 2 describes provenance for the discussion in this paper followed by surveys of some major models and standards for preservation of digital resources and provenance description. Section 3 discusses the provenance description using PROV-O and PREMIS OWL ontology respectively. Section 4 shows mapping between PROV-O and PREMIS OWL ontology and proposes a novel model to combine them for provenance description oriented to digital preservation. Section 5 states metadata schema provenance issues for metadata longevity. Finally, Section 6 concludes the paper.

2. Survey of Provenance Description Standards and Models

2.1. Digital Provenance and Metadata Provenance

We discuss provenance from the dual viewpoints of digital object provenance and that of metadata. Digital provenance and metadata provenance in this paper are defined as follows:

Digital provenance is chronology or chronological information related to management of a digital object. Digital provenance typically describes agents responsible for the custody and stewardship of digital objects, key events that occur over the course of the digital object's life cycle, and other information associated with the digital object's creation, management, and preservation (PREMIS Editorial Committee, 2012)—e.g., the organization responsible for eBook.

Based on the definition above, we can define ***metadata provenance*** as chronology or chronological information about metadata, typically responsible agents, influencing actions, associated events and other related information about metadata over its lifecycle. Provenance about metadata schema is also metadata provenance, e.g., actions and events in the revision process of metadata schema, and so forth.

It is important for memory institutions to record and provide provenance information of their holdings. W3C Provenance Incubator Group listed provenance-related use cases, which include provenance in cultural heritage (W3C Provenance Incubator Group, 2010). Europeana provides access to resources held at cultural heritage institutions throughout Europe. Europeana is a use case of metadata provenance, in which metadata provenance is represented via Europeana Data Model using the OAI-ORE model (Eckert, 2012).

The paragraphs below summarize digital provenance and metadata provenance from the viewpoint of long-term use of digital objects:

(1) Metadata of preserved resources has to be consistently interpretable over time. It has to be recognized that preservation policy and environment of preserved resources may change over time and metadata interpretation may be affected by the changes. For example, in the case of recordkeeping, digital provenance could provide information about the origin, e.g., where, when, by whom, and how a resource was created and who are the successors of the preserved resource. This information will contribute to the interpretation of metadata by users in the future.

(2) Metadata provenance describes and keeps track of responsible agents, influencing actions, associated events that caused a change(s) in metadata. Change history of a metadata schema used in a service is crucial to keeping track of changes to metadata instances created based on that schema. Therefore, provenance of a metadata schema is crucial to keeping metadata correctly and consistently interpretable and may include change history of the schema as well as relationships to other entities such as base standards and system requirements.

2.2. Digital Preservation Standards—OASIS and PREMIS

The OAIS Reference Model is a widely used model for archiving and preserving digital resources. Provenance information in OAIS is defined as the history of the Content Information, which describes the origin of and changes on an archived resource, and agents who hold custody since its origination (Consultative Committee for Space Data System, 2012). The provenance description is a part of Preservation Description Information (PDI), and documents evolutionary processing history associated with the Content Information over its complete life cycle.

PREMIS is a widely used international metadata standard for the preservation of digital objects. The PREMIS Data Model defines five *Entities* for digital preservation, which are *Intellectual Entity*, (*Digital*) *Object*, *Event*, *Agent*, and *Right*. Documentation of actions on a digital object is critical for the maintenance of the object. The documentation, i.e., metadata about the actions, is aggregated as an *Event*. Thus, *Event* is crucial component for provenance description associated with *Object*. PREMIS Data Dictionary defines a set of descriptive elements of the five *Entities*. Those elements are called semantic units. Some of the semantic units associated with an *Event* record changes to a preserved digital object (PREMIS Editorial Committee, 2012). PREMIS OWL ontology defines classes and properties to describe preservation metadata in RDF.

2.3. Provenance Models—W3C PROV, Open Provenance Model and others

W3C PROV: The Provenance Working Group at W3C has published PROV family of documents, including the PROV Data Model (PROV-DM), PROV-O and so forth. The working group aims at the inter-operable interchange of provenance information in heterogeneous environments such as the Web. PROV-DM is a conceptual data model, which defines a set of concepts and relations to represent provenance (Moreau et al., 2013). PROV-O defines a set of classes and properties as an OWL2 ontology allowing mapping PROV-DM to RDF (Lebo et al., 2013).

Open Provenance Model (OPM): OPM is a research result of the International Provenance and Annotation Workshop (IPAW). Based on the OPM Core Specification (v1.1), the OPM is designed to meet six requirements, including: exchange of provenance information between systems, representation of provenance for any "thing" and so forth (Moreau et al., 2010). OPM Vocabulary (OPMV), OPM OWL Ontology (OPMO) and OPM for Workflows (OPMW) are defined pertaining to OPM. OPMV as an OWL-DL ontology designed to assist the interoperability between provenance information on the Semantic Web and to support provenance descriptions for datasets beyond those in the Web of Data (Zhao, 2010). OPMO as an OWL ontology allows full expressivity of OPM concepts and supports inferencing (Moreau et al., 2010). OPMW is also OWL-DL ontology developed to represent abstract workflows and workflow execution traces. OPMW extends and reuses OPM's core ontologies. In the latest release, OPMW also extends PROV to represent scientific processes (Garijo and Gil, 2014).

Others: W7 model was developed o represent the semantics of data provenance in which provenance is conceptualized as a combination of seven interconnected elements including "what (occurring event)", "how (action leading to event)", "who (involved individuals or organizations)", "when (time of event)", "where (location of event)", "which (software or instrument that was used)" and "why (reason for why event happened)" (Liu, 2011). A Vocabulary for Data and Dataset Provenance (Voidp) defines terms to describe provenance relationships of data in linked datasets (Omitola et al., 2011). Provenance Vocabulary (PRV) as an OWL-DL ontology defines classes and properties for describing provenance of linked data on the Web. PRV is a domain specific specialization of PROV-O. It is notable that PRV defines terms for both data creation and data access (Hartig and Zhao, 2012). Provenance, Authoring and Versioning Ontology (PAV) is designed for the capture of essential descriptions for tracking the provenance, authoring and versioning of web resources (Ciccarese et al., 2013). BBC Provenance Ontology is designed to capture data about the provenance of data in an RDF Triple Store (BBC, 2012). Provenir Ontology (PO) defined in OWL-DL describes the classes and the properties to represent provenance metadata in eScience (Sahoo and Sheth, 2009).

2.4. Discussion on Provenance Description Standards and Models

Provenance may be about any resource, such as documents, rare books, web pages, datasets, transaction execution records, etc. This means that we need to use an appropriate vocabulary or vocabularies for provenance description in accordance with the type of resources and archiving purposes. Provenance description in OAIS and PREMIS is primarily for digital preservation whereas those standards shown in section 2.3 are defined for other purposes. Most of the ontologies are OWL-based; thus, the OWL-based definitions are useful for the reference to term definitions and reasoning of provenance.

PROV is designed generally and comprehensively for provenance description, referring to representation, interchange, query, access, and validation of provenance. PREMIS is widely used for digital preservation where provenance description is an important component. This study is primarily aimed at definition of a model of metadata provenance description for long-term use of metadata. We use PROV and PREMIS as a basis for general provenance description and provenance description for preservation in this study. Hereafter we will refer to PROV and PREMIS instead of PROV-O and PREMIS OWL Ontology unless we need to explicitly state the ontology.

3. Provenance Description Scenarios for Preservation

We use PROV-O and the PREMIS OWL Ontology to describe provenance information created during the lifecycle of digital objects and their metadata. Migration is a widely used method to assure digital objects accessible and usable over time. This section presents some instnaces of provenance description about the format migration shown below, referring to the *generationActivity*/*creationEvent* occurred to *Digital Object A*, responsible *Agent*, related date time, and also the derivation of *Digital Object A* in Format X to *Digital Object B* in Format Y via *migrationActivity* which caused the format change, and so forth.

3.1. Description of Activity and Event

Figure 1 shows a *generationActivity* leading to the generation of *Object A* by using PROV. The *generationActivity* (started at dateTime1, ended at dateTime2) resource is directed to *Object A*, which is linked to a generation Date-Time literal. PREMIS uses preservation-specific value vocabularies defined by Library of Congress. Those vocabularies provide terms expressed in SKOS vocabulary, e.g., *EventType*, *AgentType* and *RelationshipType.* Likewise, Figure 2 shows a *creationEvent* associated with *Object A* and the *creationEvent* happening during a period from dateTime1 to dateTime2. Meanwhile, the Figure also presents the *creationEvent* is linked to an *EventOutcomeInformation* resource, an *EventType* resource, and *EventDateTime* literal.

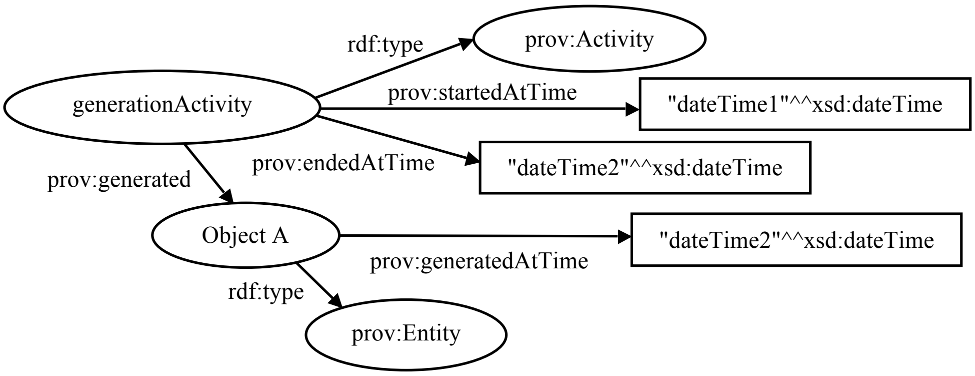


FIG.1. Provenance graph of generationActivity happened on Digital Object A using PROV

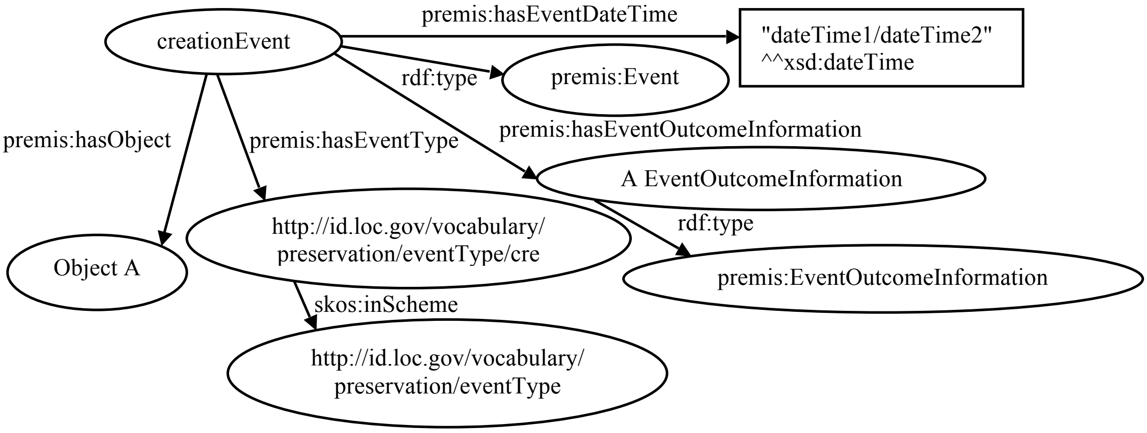


FIG.2. Provenance graph of creationEvent occurred to Digital Object A using PREMIS

3.2. Description of Responsible Agent

As shown in Figure 3, *Object A* is connected with a *Person* by property *wasAttributedTo* defined in PROV. The *generationAcitity* is linked to that *Person* via property *wasAssociatedWith*, from which we know the *Person* holds a responsibility for the generation of *Object A*. In PREMIS, *Agent* influences *Object* through *Event*. That is, *Agent* is not directly connected to *Object* as shown in Figure 4. However, PROV allows *Agent*, *Entity* and *Activity* to be related with each other directly.

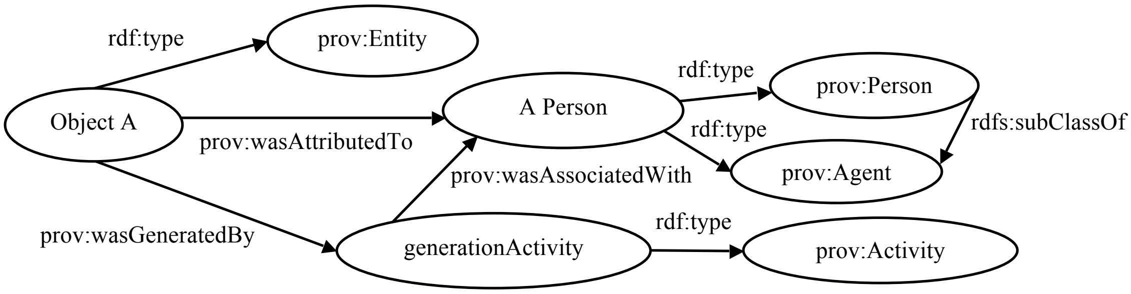


FIG.3. Provenance graph of Agent responsible for the generation of Digital Object A Using PROV

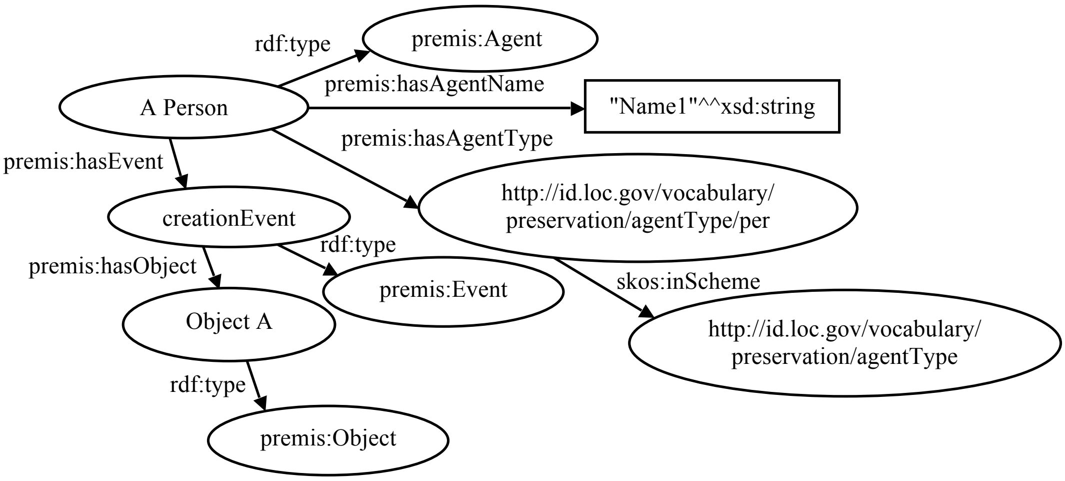


FIG.4. Provenance graph of Agent responsible for Event using PREMIS

3.3. Description of Relationships between Entities and Relationships between Objects

PROV describes the relationship between entities with the properties *wasDerivedFrom*, alternateOf, *specializationOf*, *wasQuotedFrom*, *wasRevisionOf*, *hadPrimarySource*, *hadMember*. Figure 5 shows that Object A is the primary source of Object B using PROV. PREMIS holds two types of relationship between Objects, including structural and derivation relationships defined in a SKOS vocabulary by Library of Congress. Using PREMIS, Figure 6 shows the derivation relationship between Object A and Object B due to the *migrationActivity*.

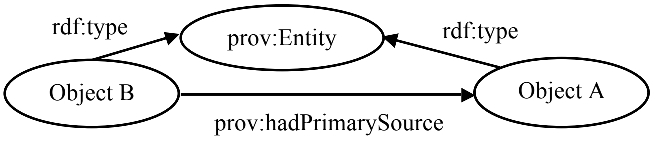


FIG.5. Derivation Relationship between Digital Object A and Digital Object B using PROV

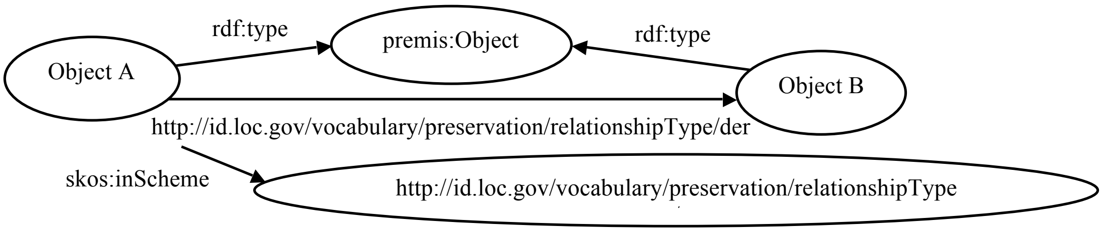

rdf:type
premis:Object
rdf:type
Object A
Object B
http://id.loc.gov/vocabulary/preservation/relationshipType/der
skos:inScheme
http://id.loc.gov/vocabulary/preservation/relationshipType

FIG.6. Derivation relationship between Digital Object A and Digital Object B using PREMIS

Furthermore, PROV also defines relationships between *Activities* and relationships between *Agents*, whereas PREMIS does not include those relationships. Figure 7 shows the relationship expressed by property *wasInformedBy* between the *migrationActivity* and *generationActivity*, which means the *migrationActivity* used *Object A* created by the *generationActivity*.

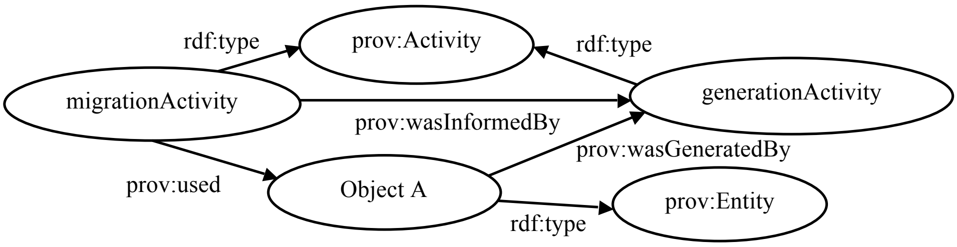


FIG.7. Relationship between Activities in PROV

4. A Merged Model for Provenance Representation by Integrating PROV-O with PREMIS OWL Ontology

4.1. Mapping of Basic Classes between PROV-O and PREMIS OWL Ontology

PROV has the three base classes, i.e., *prov:Entity*, *prov:Agent* and *prov:Activity*. PREMIS defines classes, including *premis:IntellectualEntity*, *premis:Object*, *premis:Agent*, *premis:Event*, and so forth. Based on the interpretation in PROV (Lebo et al., 2013) and PREMIS (PREMIS Editorial Committee, 2012), the paragraphs below discuss mappings between them.

premis:IntellectualEntity is a set of content items as a single intellectual unit, e.g., book, map, photograph, or database. *premis:Object* is a discrete unit of information in digital form. *prov:Entity* can be in physical or digital or conceptual or imaginary thing. We can conclude that *prov:Entity* has a broader meaning than *premis:IntellectualEntity* and *premis:Object*. Hence, we map *premis:IntellectualEntity* and *premis:Object* as subclass of *prov:Entity*.

premis:Event indicates a description about an action (or activity) impacting an *Object. prov:Activity* means actions or processes performed by *Agent(s)* or acted on *Entity (-ies)*. *premis:Event* is oriented to preservation actions, and only important *Events* are recorded. On the other hand, *prov:Activity* does not have limitation of action domain or types. That is, the meaning of *premis:Event* is narrower than *prov:Activity*. Therefore, we map *premis:Event* as subclass of *prov:Activity*.

premis:Agent can be a person, or an organization, or a software program/system associated with *Events* in the life of an *Object*. *prov:Agent* bears responsibility for occurred *Activity*, or the existence of *Entity*. However, their *Agent* types are almost the same. In a sense, *premis:Agent* can be seen to be equal to *prov:Agent*. And the relation can be described using *owl:equivalentClass*.

4.2. A Proposed Model Integrating PROV-O with PREMIS OWL Ontology

Both PROV and PREMIS have properties to describe provenance, and they are defined based on RDF and OWL. PROV is designed for generalized provenance description and interchange among different systems, whereas PREMIS is primarily for preservation metadata description used for digital preservation. The specialized PREMIS terms used to describe preservation could enrich expressive power of PROV. By introducing the controlled vocabulary for event types suggested in PREMIS, interoperability of *Activity* descriptions in PROV could be enhanced.

Based on the mapping shown in section 4.1, we propose a provenance description model for preservation of digital resources and metadata, by integrating the PROV with PREMIS. The merged model shown in Figure 8 introduces the *premis:Object* and *premis:IntellectualEntity* as the subclass of *prov:Entity*, *Collection*, *Bundle*, and *Plan* are also subclasses of *Entity*. Meanwhile, *premis:Event* is mapped to the subclass of *prov:Activity, premis:Agent* is equivalent to *prov:Agent*. In the Figure, the classes in PROV are written in italic, and the classes in PREMIS are shown with underline. Moreover, as shown in Figure 8, the relationships between classes, the generation or invalidation time of *Entity*, and the start or end time of *Activity*/*Event* can also be described via properties (written with namespace prefix, i.e., prov) from PROV.

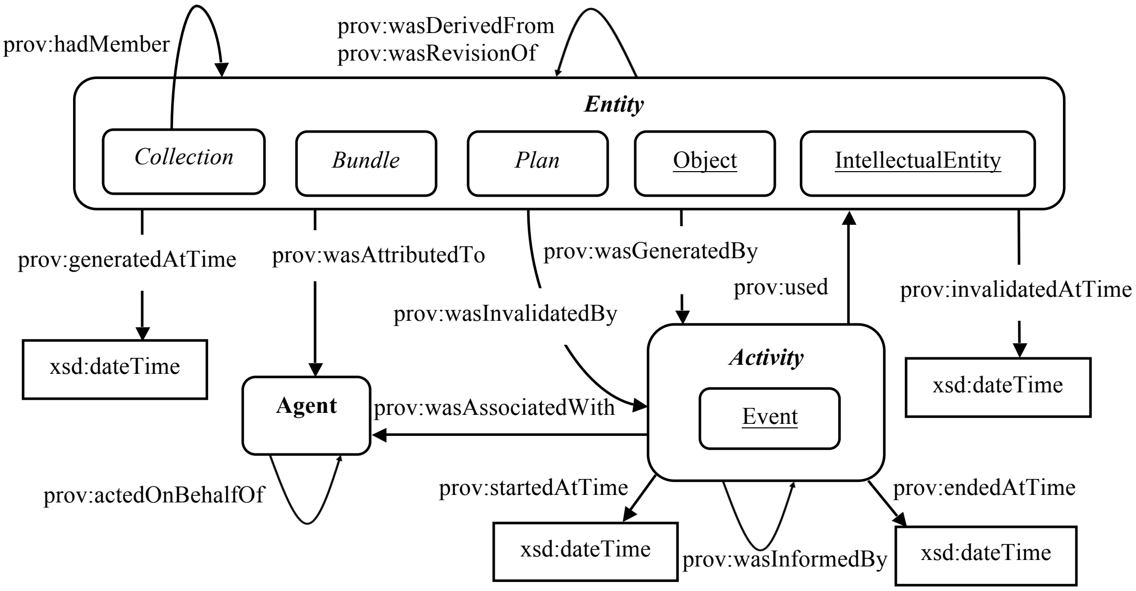


FIG.8. The merged model for provenance description oriented to digital preservation

4.3. Provenance Description Using the Proposed Model

Eckert presented the concept of Provenance Context. A Provenance Context can be seen as a Named Graph about identified resource (Eckert, 2013). Named Graph may be used for tracking provenance of RDF data, replication of RDF graphs, and versioning (Dodds and Davis, 2012). PROV allows grouping of provenance description and defines *Bundle* as a named set of descriptions (Lebo et al., 2013).

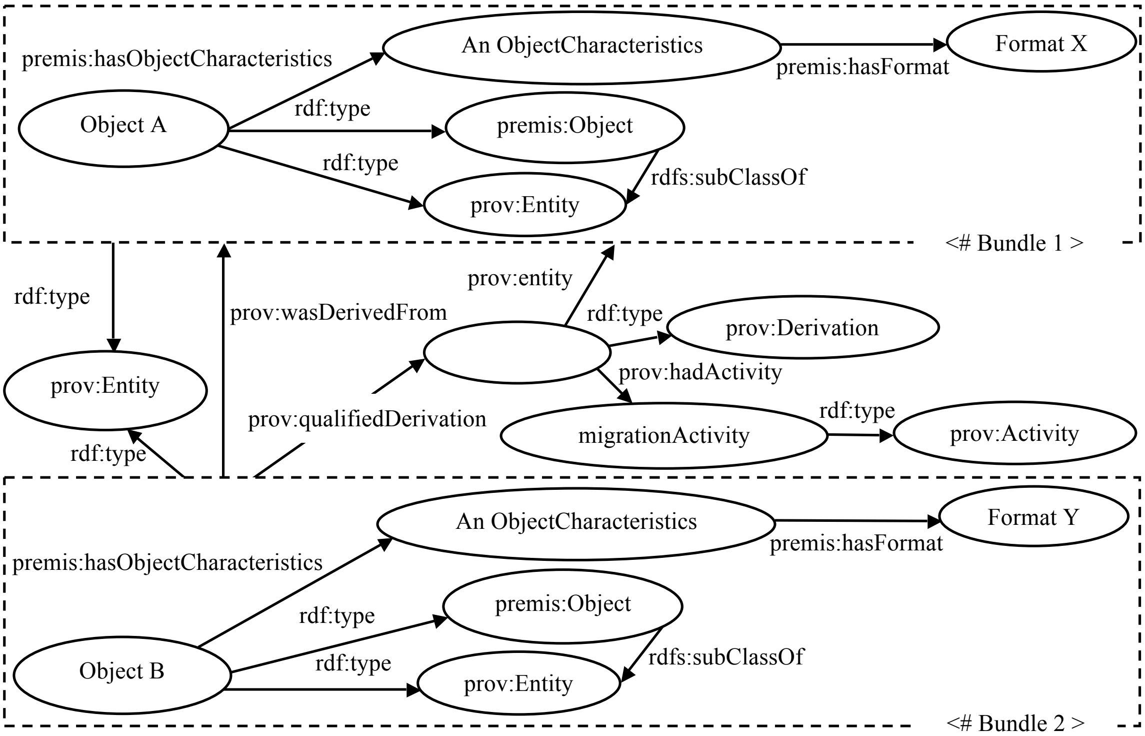


FIG.9. Provenance graph of the format change from Digital Object A to Digital Object B using Bundle

Through the definition of *Bundle*, we can describe the provenance of *Bundle*. For the assumed example, *Digital Object A* in Format X is migrated to *Digital Object B* in Format Y. Here, we define two *Bundles*, i.e., *Bundle 1* and *Bundle 2*. *Bundle 1* and *Bundle 2* respectively describes the format feature of *Digital Object A* and *Digital Object B* as shown in Figure 9, which shows the format change caused by *migrationActivity*. As *Bundle* is an *Entity* in PROV, we can also express the derivation between *Bundle 1* and *Bundle 2*. In PROV, by using property *qualifiedDerivation*, we can qualify how *Bundle 2* was derived from *Bundle 1*. In Figure 9, *Bundle 2* is linked to a blank node through property *qualifiedDerivation*. And from the blank node, the *migrationActivity* caused the format change is expressed.

5. Provenance Description for Long-term Use of Metadata

Metadata schema longevity is a vital aspect of metadata longevity. Given to the necessity of provenance in preservation, metadata schema provenance should be documented and managed with a purpose for metadata preservation. On one hand, a metadata is a digital object, and on the other hand, a metadata is a logical data entity neutral to any particular physical representation as a digital object. There are widely accepted standards for the longevity of digital objects, e.g., OAIS and PREMIS. However, there is no well-established model or standards for the longevity of metadata as a logical data entity. In this paper, the authors propose a model for provenance description of metadata from the viewpoint of metadata longevity.

By the nature of metadata, there is meta-metadata and meta-meta-metadata which mean "data about metadata" and "data about meta-metadata". Metadata schema is a typical meta-metadata because it is a description of metadata from the viewpoint of structural and/or semantic definition. Because of the nature of metadata, meta-metadata and meta-meta-metadata are metadata.

Metadata instances are created as (1) a digital instance of metadata, e.g., a text file describing a book, a CSV file of bibliographic records, or (2) a logical data instance expressed as a self-contained digital object or embedded in a digital object, e.g., a metadata expressed as an RDF/XML instance and an RDFa expression embedded in an HTML document. In both cases, provenance is an important issue for the longevity of metadata - they require both digital object provenance and metadata provenance, i.e., metadata instance as a file and a written instance in the file.

Provenance of the metadata schema is one of the key issues for the long-term use of metadata instances. Metadata schema provenance can be categorized using DCMI application profile – (1) Vocabulary Provenance, (2) Structural Provenance (i.e., provenance of description set profiles), (3) Provenance of other components: Encoding Syntax Guidelines, User Guidelines, and Functional Requirements. Vocabulary provenance is for recording semantic change of terms. Structural provenance includes revision history of terms used in the schema as well as the revision history of structural constraints. Other provenance descriptions are crucial for readers in the future to understand contextual information to process metadata. From another viewpoint, a vocabulary mapping table created for a metadata schema mapping is a metadata instance about the metadata schema mapping, e.g., conversion from an old schema to a new schema, and merger of two schemas. Provenance description for the table should be given to record a change history of metadata terms used in the schema(s).

6. Discussion and Future Work

Although many projects have made great efforts for digital preservation, there is no efficient method proposed for metadata preservation. Metadata provenance for metadata longevity in the Semantic Web is an important issue. It is easier to collect and merge open metadata from various sources. Given to the dynamic factors, e.g., URI, linkage relation, and RDF vocabulary, the representation of provenance of metadata and metadata schema is necessary.

There is a challenge in how to make metadata provenance interoperable and semantic even preservation environment changes during a long time period. Interoperability in provenance description is useful for the interchange among various domains or systems. Semantic provenance is required to make the meaning of provenance easily and correctly understandable by both humans and machines. In any event, preservation context and provenance context for metadata need further research.

References

BBC. (2012).Provenance Ontology. Retrieved March 18, 2014, from http://www.bbc.co.uk/ontologies/provenance.

Consultative Committee for Space Data System. (2012,June). CCSDS 650.0-M-2.Reference model for open archival information system (OAIS), Recommended Practice, Issue 2.Retrieved March 18, 2014, from http://public.ccsds.org/publications/archive/650x0m2.pdf.

Ciccarese, Paolo, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair JG Gray, Carole Goble, and Tim Clark. (2013).PAV 2.0 - Provenance Authoring and Versioning ontology. Journal of Biomedical Semantics 2013, 4:37. Retrieved March 18, 2014, from http://www.jbiomedsem.com/content/4/1/37.

Dodds, Leigh and Ian Davis. (2012, May 31). Chapter 5. Data Management Patterns. Linked Data Patterns: A pattern catalogue for modeling, publishing, and consuming Linked Data. Retrieved March 18, 2014, from http://patterns.dataincubator.org/book/named-graphs.html.

Eckert, Kai. (2012). Metadata Provenance in Europeana and the Semantic Web. Retrieved July 25, 2014, from http://edoc.hu-berlin.de/series/berliner-handreichungen/2012-332/PDF/332.pdf.

Eckert, Kai. (2013). Provenance and Annotations for Linked Data. Proceedings of the International Conference on Dublin Core and Metadata Applications.2013, 9-18.

Garijo, Daniel and Yolanda Gil. (2014, July 11). The OPMW-PROV Ontology. Retrieved July 29, 2014, from http://www.opmw.org/model/OPMW/.

Hartig, Olaf and Jun Zhao. (2012, March 14).Provenance Vocabulary Core Ontology Specification. Retrieved March 18, 2014, from http://trdf.sourceforge.net/provenance/ns.html.

Liu, Jun. (2011). W7 Model of Provenance and its Use in the Context of Wikipedia.Ph.D. Dissertation. The University of Arizona.

Lebo, Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao.(2013, April 30). PROV-O: The PROV Ontology. Retrieved March 18, 2014, from http://www.w3.org/TR/prov-o/.

Moreau, Luc, Paolo Missier, Khalid Belhajjame,Reza B'Far,James Cheney, Sam Coppens,Stephen Cresswell, Yolanda Gil,Paul Groth,Graham Klyne, Timothy Lebo,Jim McCusker,Simon Miles,James Myers, Satya Sahoo,and Curt Tilmes.(2013, April 30).PROV-DM: The PROV Data Model. Retrieved March 18, 2014, from http://www.w3.org/TR/prov-dm/.

Moreau, Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan and Jan Van den Bussche. (2011). The Open Provenance Model Core Specification (v1.1). Future Generation Computer Systems, 27, (6), 743-756.

Moreau, Luc, Li Ding,Joe Futrelle,Daniel Garijo Verdejo,Paul Groth,Mike Jewell,Simon Miles,Paolo Missier,Jeff Pan,and Jun Zhao. (2010, October 12). Open Provenance Model (OPM) OWL Specification. Retrieved March 18, 2014, from http://openprovenance.org/model/opmo.

Omitola, Tope, Christopher Gutteridge, and Nicholas Gibbins.(2011).voidp: A Vocabulary for Data and Dataset Provenance. Retrieved March 18, 2014, from http://www.enakting.org/provenance/voidp/.

Pearce-Moses, Richard. (2005). A Glossary of Archival and Records Terminology (pp. 317). Chicago: The Society of American Archivists. Retrieved March 18, 2014, from http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf.

PREMIS Editorial Committee. (2012). PREMIS Data Dictionary for Preservation Metadata, version 2.2. July 2012. Retrieved March 18, 2014, from http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

Sahoo, S. Satya, and Amit P. Sheth. (2009). Provenir ontology: Towards a Framework for eScience Provenance Management. Retrieved March 18, 2014, from http://corescholar.libraries.wright.edu/knoesis/80.

W3C Provenance Incubator Group.(2010).Use Case Report. Retrieved July 25, 2014, from http://www.w3.org/2005/Incubator/prov/wiki/Use_Case_Report.

Zhao, Jun.(2010, October 6).Open Provenance Model Vocabulary Specification. Retrieved March 18, 2014, from http://open-biomed.sourceforge.net/opmv/ns.html.

Interlinking Cross Language Metadata Using Heterogeneous Graphs and Wikipedia

Xiaozhong Liu
School of Informatics and Computing
Indiana University, USA
liu237@indiana.edu

Miao Chen
School of Informatics and Computing
Indiana University, USA
miaochen@indiana.edu

Jian Qin
School of Information Studies
Syracuse University, USA
jqin@syr.edu

Abstract

Cross-language metadata are essential in helping users overcome language barriers in information discovery and recommendation. The construction of cross-language vocabulary, however, is usually costly and intellectually laborious. This paper addresses these problems by proposing a Cross-Language Metadata Network (CLMN) approach, which uses Wikipedia as the intermediary for cross-language metadata linking. We conducted a proof-of-concept experiment with key metadata in two digital libraries and in two different languages without using machine translation. The experiment result is encouraging and suggests that the CLMN approach has the potential not only to interlink metadata in different languages with reasonable rate of precision and quality but also to construct cross-language metadata vocabulary. Limitations and further research are also discussed.

Keywords: metadata; linked data; cross language; heterogeneous graph

1. Research Problem

Subject categories and keywords in metadata descriptions are primary subject access points for information discovery whether for English- or non-English-speaking users. While many non-English speaking users can read and understand English, it is often not the same for the opposite. To bridge the gap between languages, digital libraries such as Europeana (http://europeana.eu) offer cross-language metadata so that users can search by any language. The cross-language search function is valuable and enables information discovery in languages that users would have otherwise unable to reach due to the language barrier.

Cross-language subject tools for Asian languages, however, have been lagging behind the increase in Asian Internet users and research output. Although the Internet has created a global village, the lack of cross-language metadata prevents information from flowing bi-directionally between English and Asian languages and creates language silos of information. Take CiNii (http://ci.nii.ac.jp/) as an example: even though both Japanese and English resources are indexed in the CiNii database, cross language retrieval and recommendation is unavailable. The same problem exists in Google Scholar, a giant scholarly retrieval engine. In addition, current tools are often limited to standardized human or machine translation, which is not suitable for high quality information retrieval and recommendation. One contributing factor for the lack of cross-language information discovery and recommendation is the difficulty in constructing a multi-language metadata vocabulary.

It is well known that the construction of any vocabulary tool is time consuming and intellectually laborious. The Chinese language version (AAT-Taiwan) of the Art and Architecture Thesaurus ("AAT", 2014) for example, is translated and mapped with its English version. It contains 34,961 concepts, 26,813 translated concepts, 12,668 archived records, and 6,564 edited records and took multiple years and professionals and domain experts to complete. The maintenance and updating has been ongoing since its first release in 2009. Building cross-language counterparts is a huge endeavor and costly in both time and personnel.

The usefulness/lack of cross-language subject vocabularies calls for new approaches to developing such vocabularies at a large scale while maintaining a reasonable level of quality and low cost. To address this conundrum, we propose a cross-language metadata network approach that will generate cross-language vocabularies on the fly by leveraging existing vocabulary resources. This paper reports a preliminary experiment as a proof of concept that uses metadata from four elements – publication, author, keyword, and venue – to construct cross-language metadata network graphs, which will then be linked through the language counterparts in Wikipedia concepts and subject categories. This approach will allow for searches in a user's native language to return results in multiple languages without machine translation.

2. Relevant Research

Developing cross-language metadata network graphs is motivated not only by the need for such tools but also by the issues in cross-language information retrieval that previous research has ignored or unable to address (Oard & Diekema 1998; Nie 2010; Ye, Huang, He & Lin 2012). Cross-language retrieval algorithms and methods are well documented in research publications. Most of these algorithms and methods, however, focused on translation rather than linking. They employed statistical models, i.e., latent semantic indexing (Littman, Dumais & Landauer 1998), parallel corpuses mining (Nie et al., 1999), and n-gram (AbdulJaleel & Larkey 2003) to construct bilingual translation models. As such, the translations rely on the source text and are limited to matching terms for translating the query from its original language to the target language in order to perform searches, rather than for linking relevant concepts cross languages. The translations have nothing to do with the metadata describing the source, much less creating both content and language linkages between metadata descriptions.

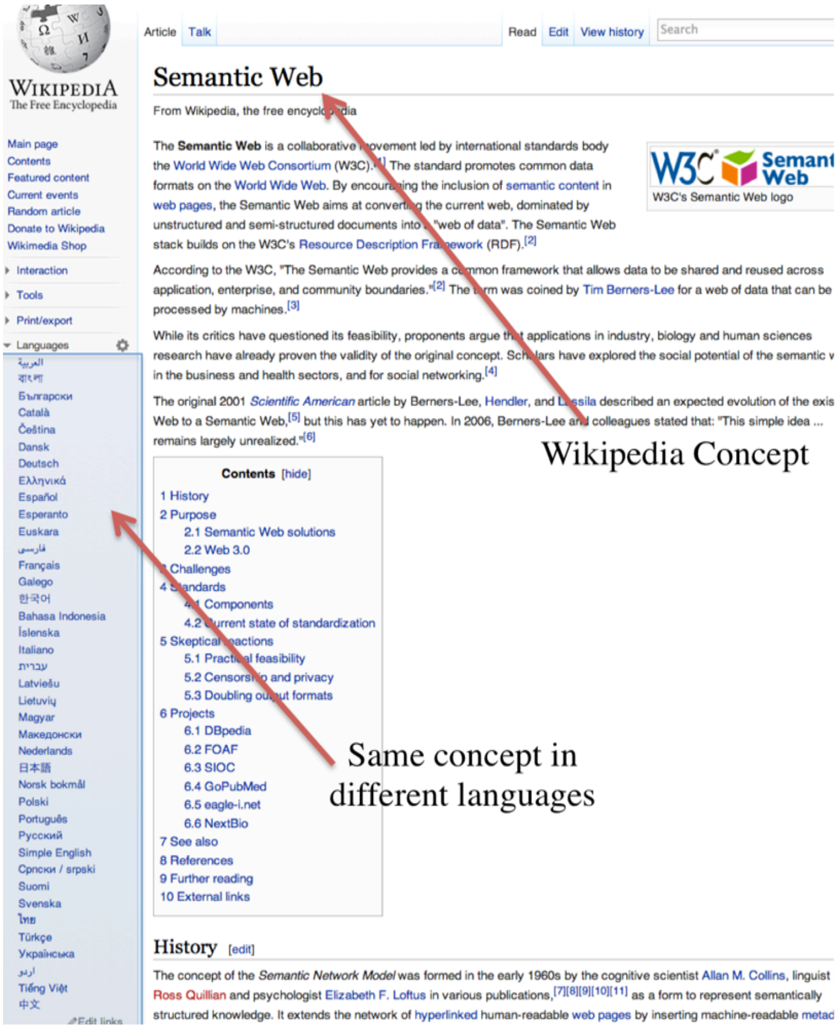
Machine translation plays an important role in constructing cross-language vocabularies (Dumais et al., 1997; Vossen 1998). Research literature in this field exhibits two paradigms of translating approaches: dictionary/rule based and parallel/comparable corpus based (Potthast, Stein, and Anderka, 2008). The first approach relies heavily on corpora and dictionaries while the second one uses the human-built cross-language links in knowledge bases such as Wikipedia. Cross-language links in Wikipedia explicitly connect concepts in different languages together and have proved to be useful sources for text mining across languages by navigating between the links. Studies show that same language pairs have a high ratio of cross-lingual links in Wikipedia. For example, the ratio of English-German links is as high as 95% (Sorg & Cimiano, 2008).

The method used by Sorg and Cimiano (2008) and Potthast et al. (2008) is called CL-ESA (Cross-Language Explicit Semantic Analysis). By projecting documents/queries to a vector space of concepts via Explicit Semantic Analysis (ESA) in one language, the vector space of concepts is mapped to a vector space of another language via cross-language links in Wikipedia. Potthast et al. (2008) used cross-language links in Wikipedia for cross-language information retrieval and showed a reasonably good performance in cross-language ranking and bi-lingual correlation ranking. Ye, Huang, He and Lin (2012) also employed Wikipedia as a graph-based bi-lingual resource for constructing a cross-language association dictionary (CLAD). They also found CLAD can be useful to enhance the cross-language information retrieval performance.

The studies mentioned above provide encouraging results for using Wikipedia as the bridge in developing cross-language metadata vocabularies. Although unforeseen factors may affect the precisions and coverage of concepts cross languages, it is nonetheless a worthwhile attempt in experimenting with the cross-language metadata linking approach using Wikipedia.

3. A Case Scenario in Cross-Language Vocabulary Linking

To demonstrate how cross-language vocabulary might be interlinked, we present a case scenario of metadata for scholarly publications. The DBLP Computer Science Bibliography (http://dblp.uni-trier.de/db/) contains metadata descriptions primarily for computer science publications written in English. The C-DBLP ("Chinese DBLP", n.d.) serves same goal for computer science publications written in Chinese. The metadata schemas for both DBLP and C-DBLP are comparable but do not communicate to one another, nor can users conduct searches across both databases. While different ownerships for each of these two databases is a primary factor for their inability to communicate to one another, it is also true that the metadata in two databases represent two completely different sets of publications and are in two different languages. Similarly, large search engine players such as Google Scholar and OCLC WorldCat index resources in multiple languages, but the metadata descriptions (e.g., keywords in different languages) in these systems are not related within their own system.



Semantic Web

From Wikipedia, the free encyclopedia

The **Semantic Web** is a collaborative movement led by international standards body the World Wide Web Consortium (W3C).[1] The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF).[2]

W3C's Semantic Web logo

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries."[2] The term was coined by Tim Berners-Lee for a web of data that can be processed by machines.[3]

While its critics have questioned its feasibility, proponents argue that applications in industry, biology and human sciences research have already proven the validity of the original concept. Scholars have explored the social potential of the semantic v in the business and health sectors, and for social networking.[4]

The original 2001 *Scientific American* article by Berners-Lee, Hendler, and Lassila described an expected evolution of the exis Web to a Semantic Web,[5] but this has yet to happen. In 2006, Berners-Lee and colleagues stated that: "This simple idea ... remains largely unrealized."[6]

Contents [hide]

History [edit]

The concept of the *Semantic Network Model* was formed in the early 1960s by the cognitive scientist Allan M. Collins, linguist Ross Quillian and psychologist Elizabeth F. Loftus in various publications,[7][8][9][10][11] as a form to represent semantically structured knowledge. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metad

Figure 1. Wikipedia concepts and language links

Over the past decade, Wikipedia has become an increasingly important resource for the world knowledge. It provides two unique features that can potentially solve the aforementioned problems for cross-language information discovery. The first feature is that Wikipedia provides concept definitions in multiple languages. An example is the concept definition for "Semantic Web": this entry has been written in 39 languages (see Figure 1). In each language, the concept name is defined by the title of the article (entry). The Chinese counterpart for this concept is defined by the title "语义网", a term used in most publications for this topic in Chinese.

The other important feature is that all concepts in Wikipedia are inter-connected via Wikipedia hierarchical categories and hyperlinked among Wikipedia pages. For instance, the concepts "Semantic Web" and "metadata" are connected via the path

[Wikipedia Concept: Semantic Web] →

[Wikipedia Category: Knowledge Engineering] →

[Wikipedia Category: Knowledge Representation] ←

[Wikipedia Concept: Metadata]

In other words, all concepts in Wikipedia are inter-connected through topic links (Wikipedia categories) and cross-language equivalents.

For the purpose of generating cross-language metadata vocabularies, the interconnectedness across multiple languages between concepts and knowledge categories in Wikipedia makes it an ideal source to leverage. If Wikipedia can be used as the intermediary vocabulary, we may be able to design algorithms to "ask" it to translate metadata between different languages. This means that digital libraries and repositories in different languages may use the intermediary tool to construct cross-language metadata vocabularies for information discovery and recommendation. It will be possible then for cross-language vocabulary tools to automatically select and recommend most relevant cross-language publications *without* having to rely on machine translation. In the cases of DBLP and C-DBLP, it is possible to use Wikipedia as the intermediary nodes to interlink publications, venues, and authors in these two digital libraries, no matter which language is used to search, via the *[Keyword] →[Wikipedia Concept]* link. As each Wikipedia concept is written in both Chinese and English, this step does not need to involve machine translation.

We are aware of the limitation of Wikipedia resource, and the sparseness of Wikipedia definitions in certain languages may limit the generalizability of the proposed method. For instance, if there is only a small amount of Wikipedia concepts defined in a language, the keyword projection performance can be understandably low.

4. Methodology

Using Wikipedia to create Cross-Language Metadata Networks (CLMN) involves two steps. In the first step a Single-Language Metadata Network (SLMN) is built for a monolingual digital library or repository. In the second step, the SLMN will be mapped to Wikipedia concepts and subject categories to create Cross-Language Metadata Networks (CLMN). Through this two-step method, cross-language metadata vocabularies are constructed and then used to connect metadata and resource objects in digital libraries/repositories across different languages. In the section below we will first discuss the method for generating metadata networks for an individual repository and then describe the CLMN through which SLMNs are interconnected via Wikipedia's bridge nodes, i.e., Wikipedia pages and subject categories.

4.1 Step 1: Creating Single-Language Metadata Networks (SLMN)

We assume that there are four types of resource objects – publications (papers, reports, webpages, and books), venues (journals, conference proceedings, and domain names as embodied by websites), subjects, and authors – in a single-language digital library. Between the four types of resource objects, there exist various types of linkages: citation linkages between publications, authorship linkages between authors and publications, and venue linkages between publications and venues. We also assume that in a single-language digital library (or repository), a list of subject terms and values (keywords or controlled vocabulary) is available to represent publications and venues and that metadata and publications share the same language.

Using the network theory, each resource object is a network node (vertex) and the links between nodes (vertices) are edges. Metadata in a single-language digital library are considered as a single-language metadata network (SLMN) in which the nodes are connected by edges. This network is heterogeneous by nature in the sense of network node types, because the same network contains multi-types of nodes: author (A), publication (P), venue (V), and keyword (K), which are what this study focuses.

For each digital library (a commercial database or an institutional repository), there exists a local SLMN. All four types of nodes mentioned above can be connected by any of the 7 types of edges: 1) $P \rightarrow A$, a paper is written by an author; 2) $P \rightarrow V$, a paper is published in a venue; 3) $P \rightarrow K$, a paper or publication is relevant to a keyword; 4) $P \rightarrow P$, a publication cites or links to publications; 5) $K \rightarrow P$, a keyword (topic) is assigned to publications; 6) $K \rightarrow A$, a keyword (topic) is assigned by authors; and 7) $K \rightarrow V$, a keyword (topic) is assigned to venues. Edge types 1, 2, 3 and 4 are implemented by using metadata in a single-language digital library. Keywords derived from publications, author names, and venues are labeled as topic and represented by edge types 5, 6, and 7, which are calculated by using PageRank with Prior algorithms (Liu, Zhang, and Guo, 2013) on the homogeneous citation graphs (publication-citation graph, author-citation graph, and venue-citation graph). Note that, as this network can be potentially used for resource recommendation, all edges are associated with an edge weight, *P(v|u)*, which indicates the transitioning probability (weight) from node *u* to node *v*.

4.2 Step 2: Creating Cross-Language Metadata Networks (CLMN)

The goal of this step is to generate cross-language metadata networks using computational methods. The CLMNs generated from using Wikipedia and the PageRank Prior algorithms will function as a linking mechanism to interconnect metadata silos of single language into a global network with the capability of performing cross-language information discovery and recommendation. In the CLMN approach, a collection of digital libraries or repositories are represented by *k* Metadata Networks (MNs). Figure 2 visualizes the CLMN creation progress. There are four layers in a CLMN and *k* SLMNs connect to the Wikipedia concept and Wikipedia category nodes on the CLMN, in which Wikipedia nodes function as the bridge to interconnect different SLMNs. Meanwhile, all Wikipedia nodes (Wikipedia concepts and Wikipedia categories) also connect with the incoming/outgoing links (between Wikipedia concepts), concept-category relations, and the hierarchical relations between categories.

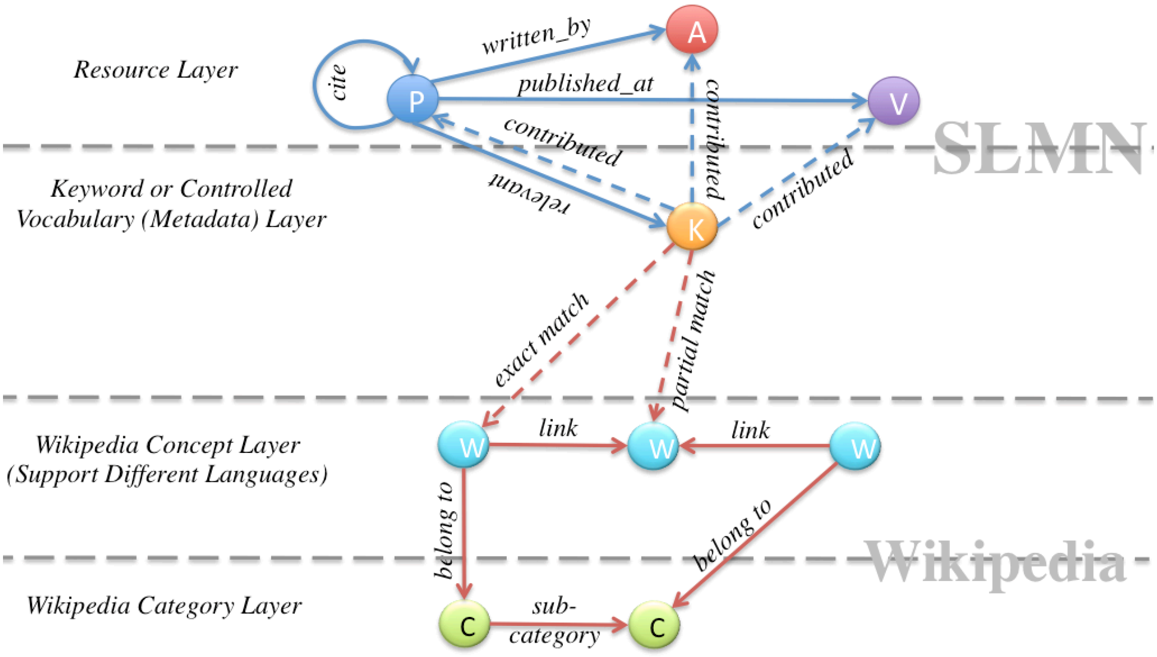


Figure 2. Cross-Language Metadata Networks (CLMN)

In Figure 2, dotted lines indicate the calculated or inferred relationships and the solid lines indicate the relationships physically exist in the repository or Wikipedia database. It depicts how one SLMN typically connects to Wikipedia nodes, which is also how other SLMNs will connect to the Wikipedia nodes. The middle section is where automatic pairing and linking of the concepts in different languages takes place. All keywords or controlled vocabularies (node *K)* connect to Wikipedia concepts via two kinds of edges: exact match edge and partial match edge. The former edge type indicates that the string represented by node *K* is exactly the same as Wikipedia concept title. Note that *K* on different SLMNs may be in different languages, while Wikipedia concept is also indexed by multiple languages. The latter edge type is generated by using information retrieval algorithms, e.g., language model or vector space model, which means that the target keyword or controlled vocabulary is part of the content of the Wikipedia concept's content. Similarly, the content of Wikipedia concept may also be in different languages. Similar as the edge types in SLMN, all edges between Wikipedia nodes and keywords nodes are associated with the edge weight.

5. Preliminary Experiment

As a proof of concept for the proposed method, we construct a CLMN by using the ACM Digital Library (English computer science publications + metadata, http://www.acm.org/) and WanFang Digital Library (Chinese computer science publications + metadata, http://www.wanfangdata.com.cn/). All four types of nodes in publications’ metadata across both libraries – authors, venues, papers, and keywords – were connected by using the intermediary layer Wikipedia as shown in Figure 2. For this experiment, we used Wikipedia Chinese and English 2014 April dumps.

Due to the space limit, we present only the metadata layer and Wikipedia layer in this section. The CLMN constructed in this preliminary experiment contains 1,481 English keywords and 121 Chinese keywords (English keywords 10 times more than Chinese keywords because of the data limitation). Connected to these keywords were 1,719 Wikipedia page nodes and 1,146 Wikipedia category nodes.

Two exemplar Chinese keywords, “机器学习” (Machine Learning) and “信息抽取” (Information Extraction) , were used as query terms to find the related English keywords by using two types of paths: 1. *[Chinese Keyword] → [Wikipedia Concept] ←[English Keyword],* and 2. *[Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword]* (Edge direction was ignored). The first path used only one intermediary Wikipedia node between the query and target keywords in Chinese and English. The second one was more complicated because the Chinese query keyword and English target keyword may link to different Wikipedia concepts and these Wikipedia concepts may share the same Wikipedia category.

Given the space limitation, we investigated only the first example in more details. Figure 3 displays the paths through which the results for “机器学习” were generated. Different types of nodes are represented by different colors on the CLMN graph. This graph example shows that Wikipedia page and category nodes function as intermediary nodes to link together the same concept Machine learning in English and Chinese.

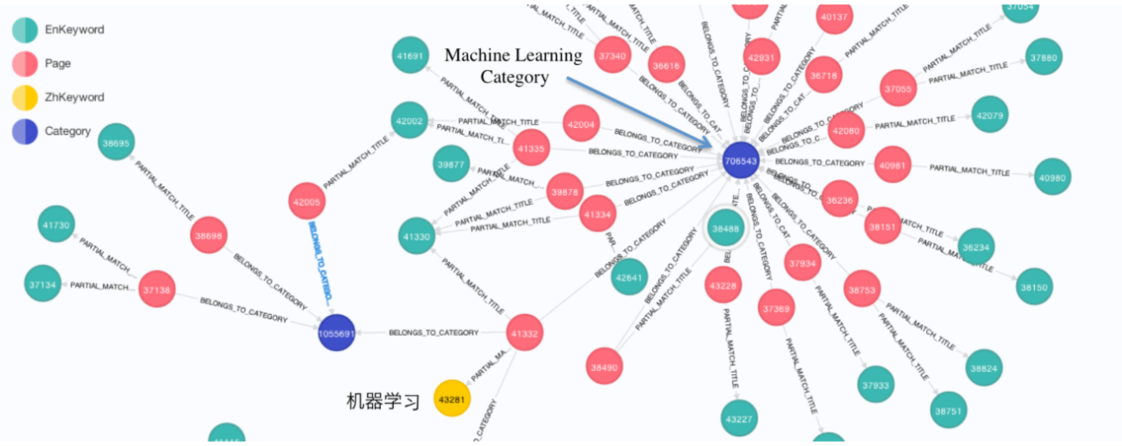


Figure 3. Related English Keywords for “机器学习” on the CLMN (via Wikipedia nodes)

The specific paths for query “机器学习” on the CLMN are listed below (CK = Chinese keyword, WP = Wikipedia page, WC = Wikipedia category, and EK = English Keyword):

Result for path [Chinese Keyword] → [Wikipedia Concept] ←[English Keyword] (1 result)

1. CK:机器学习→WP:machine_learning←EK:machine_learning

Results for path [Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword] (26 results)

1. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:cluster_analysis←EK:cluster_analysis
2. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:expectation_maximization_algorithm ←EK:em_algorithm
3. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:complex_systems←EK:complex_systems
4. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:reinforcement_learning←EK:reinforcement_learning
5. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:pattern_recognition←EK:pattern_recognition
6. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:formal_concept_analysis←EK:concept_analysis
7. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:unsupervised_learning←EK:unsupervised_learning
8. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:hidden_markov_model←EK:hidden_markov_model
9. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:expectation_maximization_algorithm ←EK:expectation_maximization
10. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:supervised_learning←EK:supervised_learning
11. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:pattern_recognition←EK:pattern_detection
12. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:artificial_neural_network←EK:neural_networks
13. CK:机器学习→WP:machine_learning→WC:Machine_learning←WP:artificial_neural_network←EK:artificial_neural_network
14. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:genetic_algorithm←EK:genetic_algorithm
15. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:nearest_neighbor_search←EK:nearest_neighbor_search

16. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:principal_component_analysis←EK: principal_component_analysis
17. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:artificial_intelligence←EK:artificial_intelligence
18. CK:机器学习→WP:machine_learning→WC:Cybernetics←WP:system←EK:systems
19. CK:机器学习→WP:machine_learning→WC:Cybernetics←WP:autonomy←EK:autonomy
20. CK:机器学习 →WP:machine_learning→WC:Cybernetics←WP:control_theory←EK:control_theory
21. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:support_vector_machine←EK:support_vector_machine
22. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:information_theory←EK:information_theory
23. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:discriminative_model←EK:discriminative_model
24. CK:机器学习 →WP:machine_learning→WC:Machine_learning←WP:perceptron←EK:perceptron
25. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:formal_concept_analysis←EK:formal_concept_analysis
26. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:conditional_random_field←EK:conditional_random_field

Specific paths for query "信息抽取" are listed below:

Results for path [Chinese Keyword] → [Wikipedia Concept] ←[English Keyword] (1 result):

1. CK:信息抽取→WP:information_extraction←EK:information_extraction

Results for path [Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword] (13 results):

1. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:artificial_intelligence←EK:artificial_intelligence
2. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:computer_vision←EK:computer_vision
3. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:description_logic←EK:description_logics
4. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:fuzzy_logic←EK:fuzzy_logic

5. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:game_theory←EK:game_theory
6. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:intelligent_agent←EK:intelligent_agent
7. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:markov_random_field←EK: markov_random_field
8. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:cross-language_information_retrieval←EK:cross_language_information_retrieval
9. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:information_retrieval←EK:information_retrieval
10. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:latent_semantic_analysis←EK:latent_semantic_analysis
11. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:natural_language_processing←EK:natural_language_processing
12. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:natural_language←EK:natural_language
13. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:question_answering ←EK:question_answering

The specific results shown above demonstrate that the path *[Chinese Keyword] → [Wikipedia Concept] ←[English Keyword]* can find accurate translation, while the path *[Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword]* can locate a number of high quality related (linked) keywords in a different language. The experiment results suggest that CLMN is promising as a means to link metadata across languages and digital libraries. The metadata used in this experiment are relatively specialized with reasonable level of quality, hence whether the method can be applied to other domains and accomplish a comparable level of performance will need further study and evaluation.

6. Discussion and Conclusion

The resulting CLMNs have a number of potentials for metadata representation and resource discovery. The four sets of results presented in the last section are structured data with path and node information attached. They can be parsed into the format suitable for building cross-language vocabularies using computer programs. Such cross-language vocabularies can be then encoded in the Linked Data formats and shared through vocabulary services. Another application is to recommend resources (i.e., publication, author or venue) across repositories and languages. For example, given an author ID (on a SLMN), the system can recommend publications potentially relevant to users' interest in a different language. Given a keyword (on a SLMN), we can recommend top related venues (venue recommendation) or expert (author recommendation) in a different language.

Unlike classical machine translation methods that use homogeneous data sources, this study employed heterogeneous graph mining and text mining methods to connect all the metadata via

Cross-Language Metadata Networks (CLMN), in which Wikipedia is used as the intermediary nodes to link local repositories. We took metadata from ACM and WanFang digital libraries to run our experiment. The results suggest that CLMN as a novel approach was able to find not only accurate translations but also locate related metadata in different languages. This is especially encouraging for developing a low cost and effective method for automatic cross-language vocabulary construction.

The reliability and validity of CLMN method need further study and experiment to verify. We plan to conduct further experiment with other sources of metadata, e.g., those available in open repositories where metadata are crowd-sourced and in disciplines other than computer science. As our next step research, we are keen on developing a bilingual vocabulary linked data set using this method in a humanities domain by leveraging data from public digital libraries.

References

AAT (ART & Architectural Thesaurus). Retrieved Aug 1, 2014 from http://www.getty.edu/research/tools/vocabularies/lod/

AbdulJaleel, Nasreen, and Leah S. Larkey. (2003). Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 139-146). ACM.

Chinese DBLP. Retrieved Aug 1, 2014 from http://cdblp.cn/index.php

Dumais, Susan T., Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. (1997) Automatic cross-language retrieval using latent semantic indexing. In AAAI spring symposium on cross-language text and speech retrieval (Vol. 15, p. 21).

Littman, Michael L., Susan T. Dumais, and Thomas K. Landauer.(1998). Automatic cross-language information retrieval using latent semantic indexing. In Cross-language information retrieval (pp. 51-62). Springer US.

Nie, Jian-Yun, Michel Simard, Pierre Isabelle, and Richard Durand. (1999, August). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval(pp. 74-81). ACM.

Nguyen, D., A. Overwijk, C. Hauff, D.R. Trieschnigg, D. Hiemstra, and F. De Jong, (2009). WikiTranslate: query translation for cross-lingual information retrieval using only Wikipedia. In Evaluating Systems for Multilingual and Multimodal Information Access (pp. 58-65). Springer Berlin Heidelberg.

Nie, Jian-Yun. (2010). Cross-language information retrieval. Synthesis Lectures on Human Language Technologies, 3(1), 1-125.

Potthast, Martin, Benno Stein, and Maik Anderka. (2008). A Wikipedia-based multilingual retrieval model. In Advances in Information Retrieval (pp. 522-530). Springer Berlin Heidelberg.

Sorg, Philipp, and Philipp Cimiano. (2008a). Cross-lingual information retrieval with explicit semantic analysis. In Working Notes for the CLEF 2008 Workshop.

Sorg, Philipp, and Philipp Cimiano. (2008b). Enriching the crosslingual link structure of Wikipedia-a classification-based approach. In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (pp. 49-54).

Vossen, Piek. (1998). A multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht.

Ye, Zheng., Huang, Jimmy X., He, Ben, and Hongfei Lin (2012). Mining a multilingual association dictionary from Wikipedia for cross-language information retrieval. Journal of the American Society for Information Science and Technology, 63(12), 2474-2487.

Automated Enhancement of Controlled Vocabularies: Upgrading Legacy Metadata in CONTENTdm

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

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

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

Abstract

To ensure robust, reliable, retrievable and sharable metadata, the University of Houston (UH) Libraries initiated a Metadata Upgrade Project in 2013 to systematically audit and refine the quality of the metadata in the University of Houston Digital Library (UHDL). Still in progress, the Metadata Upgrade Project has already produced significant discoverability improvements in the UHDL's legacy metadata and laid the foundation for future metadata production according to recognized standards. The final phase of the project includes aligning controlled vocabulary terms with appropriate authorities and adding and revising descriptive content in the UHDL. This is a time intensive process that requires careful evaluation and entry of name and subject authority terms. To improve efficiency and accuracy during the data entry process, the metadata librarian at the UH Libraries developed name and subject authority applications that automatically transform legacy controlled vocabulary terms into authorized forms. This project report provides an overview of the UH Libraries Metadata Upgrade Project, a discussion of how the UHDL's upgraded metadata improves discoverability of our collections, and an in-depth look at the custom tools that automate the authority alignment process in the CONTENTdm Project Client.

Keywords: metadata; controlled vocabularies; authority control; automation

1. Introduction

The University of Houston (UH) Libraries are committed to the dissemination and discoverability of our unique, historical collections. In the five years since the launch of the University of Houston Digital Library (UHDL), the repository has grown steadily and currently provides online access to more than 50,000 digital objects. While the UHDL serves as a platform for researchers to access the rare and unique materials in the UH Libraries holdings, the state of the legacy metadata in the digital library presented barriers to efficient use of the UHDL's digital objects. Incomplete and inconsistent legacy metadata restrict both discoverability and interoperability. To ensure robust, reliable, retrievable and sharable metadata, the UH Libraries initiated a Metadata Upgrade Project in 2013 to audit and refine the quality of the metadata in the UHDL.

The Metadata Upgrade Project team developed a three-phase strategy to systematically manage the metadata audit and upgrade process based on feedback and data analysis from focus group interviews, data inspection and benchmarking. Still in progress, the Metadata Upgrade Project has already produced significant discoverability improvements in the UHDL's legacy metadata. The third phase requires time intensive work on item level descriptive metadata revision including aligning controlled vocabulary terms with appropriate authorities. To improve efficiency and accuracy during the data entry process, the metadata librarian at the UH Libraries developed name and subject authority applications that automatically transform legacy controlled vocabulary terms into authorized forms.

2. Metadata Upgrade Methodology and Strategy

The Metadata Upgrade Project utilized several approaches to identify metadata issues and create strategies to improve the quality of metadata in the repository. To understand metadata needs and address concerns that developed around legacy metadata, librarians conducted focus groups with UH Libraries stakeholders—including Special Collections, Web Services, and Liaison Services. External stakeholders were not included in the focus group interviews because of the complicated institutional review board (IRB) application requirements and the difficulty in identifying users. The project team also benchmarked current practices with similar digital libraries. These two activities demonstrated that controlled vocabularies in the UHDL had been applied inconsistently and inaccurately over time, most likely as a result of frequent changes in staff from project to project. Consequently, some items in the UHDL had rich descriptive connections with items in different digital collections while others had no terms to link them to similar materials. The Metadata Upgrade Project team concluded that the controlled vocabulary terms in the UHDL should be revised for accuracy, standardized to specific vocabulary lists, and mapped to appropriate Dublin Core elements (Thompson and Wu, 2013).

TABLE 1: Three Phases of the Metadata Upgrade Project

Project Phase	Tasks
Phase One	Stakeholder Interviews, Metadata Schema Development
Phase Two	Collection-level Metadata Editing, Metadata Dictionary
Phase Three	Item-level Metadata Editing

After collecting data regarding the issues with the legacy metadata in the UHDL, librarians developed key recommendations, a three-phase strategy for upgrading UHDL metadata (Table 1), and a new input standard to ensure that the quality of future metadata remains accurate and consistent over time. The first phase of the upgrade process focused on adding, revising, and standardizing descriptive and administrative fields. The second phase edited metadata at the collection level. Tasks performed in phase two included standardizing collection names for archival and digital collections as well as editing collection-level fields. The third phase focuses on adding and revising descriptive content in the digital library at the item level. To ensure that future UHDL metadata complies with the new standard, the Metadata Upgrade Project also produced a Metadata Dictionary which provides definitions, examples, and input rules for descriptive, administrative, technical, and preservation metadata fields (Thompson and Wu, 2013). An abridged version of the UHDL Metadata Dictionary (2014) is available online.

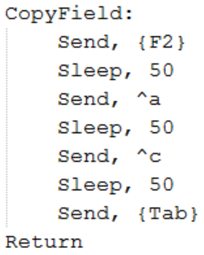
3. Automated Metadata Transformation

Addressing issues with controlled vocabulary terms is a key activity in the third phase, and the Metadata Upgrade Project staff spends a considerable amount of time reviewing existing terms, identifying more appropriate terms, and reconciling terms with the source vocabularies. In the early stages of phase three, the Metadata Upgrade Project staff experimented with exporting data from CONTENTdm and cleaning the data with OpenRefine. However, getting the cleaned data back into the system with a batch process proved a difficult task. The staff chose to work in the CONTENTdm Project Client for all phase three item-level editing and use OpenRefine for metadata analysis on new collections.

In order to speed up the editing process, the UH Libraries metadata librarian developed two applications that enable efficient transformation of legacy authority data within the CONTENTdm Project Client. Both applications are written in AutoHotkey (AHK), an open source scripting and macro language for the Windows operating system. In addition to a GUI that provides user feedback and menu functions, the core AHK scripts act as a glue language that connects the data in the Project Client with locally maintained vocabulary mapping files. Each AHK authority app gathers data recorded in the CONTENTdm Project Client and parses the tab-delimited authority files for matching terms. As of this writing, the tab-delimited files contain approximately 900 subject mappings and 3,000 name authority mappings. The apps automatically enter authorized terms in the Project Client and facilitate the addition of new terms to the local mapping files with input boxes and automatic Web browser searches. Most importantly, the apps allow the Metadata Upgrade Project team to focus on the intellectual content of their authority work and let the computer take care of repetitive data entry tasks.

3.1 Subject Authority App

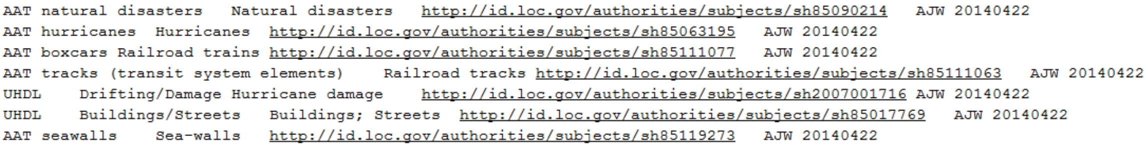
The decision to develop a subject authority app stems from the desire to ensure that the metadata for every object in the UHDL contains subject terms from a widely used controlled vocabulary. Legacy subject data in the UHDL includes terms from multiple vocabularies, and the subject app performs automated mapping from those vocabularies to authorized terms in the Library of Congress Subject Headings (LCSH). The UH Libraries are exploring opportunities for applying linked data technologies to the collections in the UHDL, and the subject app also facilitates harvesting of URIs from the Library of Congress Linked Data Service in preparation for that work.



```
CopyField:
    Send, {F2}
    Sleep, 50
    Send, ^a
    Sleep, 50
    Send, ^c
    Sleep, 50
    Send, {Tab}
Return
```

FIG. 1. AHK sub-routine for copying data and moving between Project Client fields.

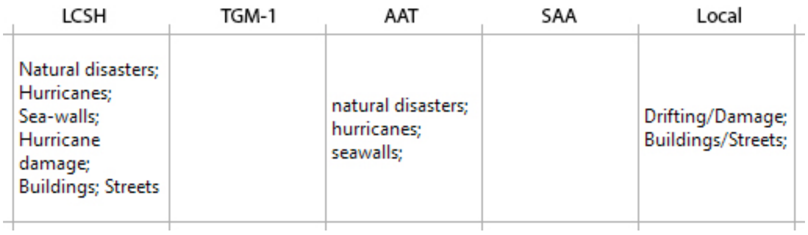
The subject authority app processes one record at a time in the Project Client's spreadsheet view. When a metadata specialist triggers the subject app with the specified key combination, the app traverses one row and copies the data in each alternate subject authority field to the clipboard. In addition to LCSH, the UHDL uses four other subject vocabularies: Thesaurus for Graphic Materials (TGM), Art & Architecture Thesaurus (AAT), the Thesaurus for Use in College and University Archives (SAA), and a local UHDL vocabulary. To move between fields and copy the data, the app sends key presses to the Project Client, as if a human user were pressing keys on the keyboard. The sub-routine in Figure 1 sends the F2 key to activate the Project Client field for editing, Control + A (^a) to select all of the text, Control + C (^c) to copy the text to the clipboard, and the Tab key to move to the right one field. Brief pauses in between each keystroke (Sleep, 50) give the Project Client GUI time to process each command.



```
AAT natural disasters	Natural disasters	http://id.loc.gov/authorities/subjects/sh85090214	AJW 20140422
AAT hurricanes	Hurricanes	http://id.loc.gov/authorities/subjects/sh85063195	AJW 20140422
AAT boxcars Railroad trains http://id.loc.gov/authorities/subjects/sh85111077	AJW 20140422
AAT tracks (transit system elements)	Railroad tracks http://id.loc.gov/authorities/subjects/sh85111063	AJW 20140422
UHDL	Drifting/Damage Hurricane damage	http://id.loc.gov/authorities/subjects/sh2007001716 AJW 20140422
UHDL	Buildings/Streets	Buildings; Streets	http://id.loc.gov/authorities/subjects/sh85017769	AJW 20140422
AAT seawalls	Sea-walls	http://id.loc.gov/authorities/subjects/sh85119273	AJW 20140422
```

FIG. 2. Subject mapping entries in the local tab-delimited file.

After copying values in a field, the app parses the clipboard data and attempts to match each term against a tab-delimited mapping file stored on a local network drive (Figure 2). If no match is found for a given term, the app opens a Library of Congress Linked Data Service search for that term in a Web browser. After identifying an appropriate controlled term, a metadata specialist enters the authorized form and authority record URI in dialog boxes. The app automatically adds the term and URI to the local mapping file. When all of the alternative subject authority columns have been queried, the app returns to the LCSH column and inputs the authorized LCSH terms for that record (Figure 3) (Weidner, UHDL_SubjectTopical_CDM, 2014).



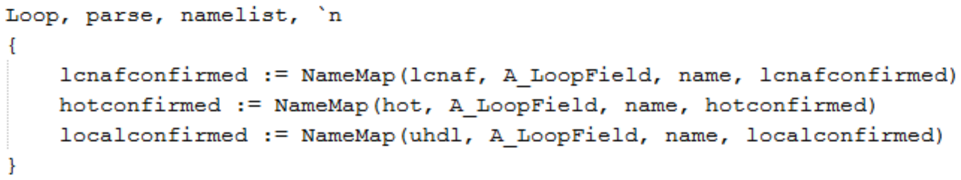
LCSH	TGM-1	AAT	SAA	Local

Natural disasters;
Hurricanes;
Sea-walls;
Hurricane damage;
Buildings; Streets
		natural disasters;
		hurricanes;
		seawalls;
				Drifting/Damage;
				Buildings/Streets;

FIG. 3. Subject values in the Project Client after mapping.

3.2 Name Authority App

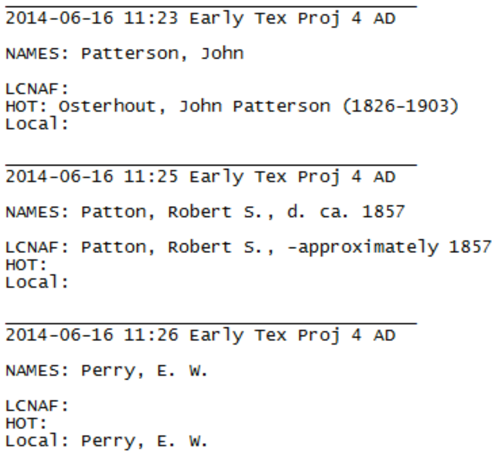
The UHDL name authority app performs similar matching and mapping functions in a different direction. Instead of mapping values in multiple columns to a single vocabulary, the name app maps values in a single column to multiple vocabularies: Library of Congress Name Authority File (LCNAF), the Handbook of Texas (HOT), and a local UHDL name authority file (Figure 4). Much of the legacy name authority data in the UHDL is recorded in the LCNAF field, even though many of those names do not have records in the LCNAF vocabulary. This occurred as a result of the metadata schema work in phase two of the Metadata Upgrade Project when staff divided the UHDL's name fields (Creator, Subject.Name, etc.) into multiple vocabularies instead of one general field. In an effort to produce high quality, standardized data that is compatible with linked data principles, the name authority app automates the transfer of name data to the appropriate authority column in the CONTENTdm Project Client (Weidner, UHDL_Names_CDM, 2014).



```
Loop, parse, namelist, `n
{
    lcnafconfirmed := NameMap(lcnaf, A_LoopField, name, lcnafconfirmed)
    hotconfirmed := NameMap(hot, A_LoopField, name, hotconfirmed)
    localconfirmed := NameMap(uhdl, A_LoopField, name, localconfirmed)
}
```

FIG. 4. AHK loop passes each name to the NameMap function which returns an authorized form.

Monitoring accuracy during authority work is very important, and the Metadata Upgrade Project staff periodically review the name app's tab-delimited mapping file in OpenRefine to identify names mistakenly mapped to more than one form. Faceting on the authorized form column quickly reveals any problems with the data. As a quality control feature, the name authority app creates a report for each day and a log entry each time the name app is triggered (Figure 5). Using these reports, staff can backtrack to locate any records that must be corrected.



```
______________________________________
2014-06-16 11:23 Early Tex Proj 4 AD

NAMES: Patterson, John

LCNAF:
HOT: Osterhout, John Patterson (1826-1903)
Local:

______________________________________
2014-06-16 11:25 Early Tex Proj 4 AD

NAMES: Patton, Robert S., d. ca. 1857

LCNAF: Patton, Robert S., -approximately 1857
HOT:
Local:

______________________________________
2014-06-16 11:26 Early Tex Proj 4 AD

NAMES: Perry, E. W.

LCNAF:
HOT:
Local: Perry, E. W.
```

FIG. 5. Name app report illustrating correct mappings to authorized forms.

3.3 Authority App Limitations

During the course of the authority work with the name and subject applications, the Metadata Upgrade Project team has identified a number of limitations. The apps can handle the bulk of the work, but there are edge cases that present interesting problems. In the case of the subject authorities, mappings to LCSH may change between collections because a single term in an alternate vocabulary can map to the multiple LCSH authorized terms. For example, the term "gutters" in an alternate vocabulary could map to "Roof gutters" or "Street gutters" in LCSH, depending on the context of the collection. This problem requires careful evaluation of a record each time the app is triggered and occasional editing of the tab-delimited subject mapping file.

In the case of the name authorities, there are many times when a name is present in both the LCNAF and HOT vocabularies. An update to the app provided the ability to harvest URIs from both vocabularies and record those connections in a separate file for future use. The app gives precedence to LCNAF for data entry purposes. As previously mentioned, the local tab-delimited name mapping file requires constant monitoring to ensure the accuracy of the authorized forms entered in the UHDL's metadata. Both AHK authority apps are short term solutions for the Metadata Upgrade Project and must eventually be supplanted by more robust controlled vocabulary management features in the UHDL's digital asset management system.

4. Benefits of Enhanced Metadata

There are numerous benefits to upgrading the legacy metadata in the UHDL. Integrating metadata best practices—including the consistent use of established controlled vocabularies—shaped the strategies and standards developed to address the issues identified during focus group interviews and benchmarking. These best practices will improve how users connect with UHDL content. In particular, standardized vocabulary terms consistently applied improve recall during faceted browsing, reducing the likelihood of orphaned records. Implementing best practices also ensures that UHDL metadata is fully interoperable with harvesting protocols, such as OAI-PMH, thereby providing another potential discovery layer to our content and opening up possibilities for collaboration with larger projects.

Aligning controlled vocabulary terms with recognized authorities and harvesting authority record URIs also lays the foundation for publishing UHDL collections as linked data with rich semantic markup. A first step might be to enrich subject terms and names with an owl:sameAs link, populated by the URI gathered during the Metadata Upgrade Project, that points to the unambiguous definition in the source vocabulary (W3C, 2004). Finally, with the creation of a more robust metadata dictionary, UHDL metadata creators now have a standard to guide future projects (Thompson and Wu, 2013).

5. Conclusion

While it is crucial to employ standards and best practices for quality control during the creation of a repository's metadata, metadata must be constantly maintained to reflect changes in the data model, end-user interface configuration, and system transitions. The lack of batch processing and limited authority control features in our digital asset management system creates barriers in our metadata editing workflow. The rapidly growing volume and complexity of formats in our digital library also presents challenges for our data quality management work. The utilization of scripting and automation in our metadata revision process has assisted us greatly in overcoming these barriers and challenges. The subject and name authority applications described in this paper have simplified our workflow and helped to improve consistency and accuracy in our data.

Metadata is at the functional core of our digital system. High quality metadata not only enhances the user experience in our digital library, but also enables the scalability and interoperability of our data. To ensure high quality metadata, it is important for metadata professionals to leverage traditional skills and new technologies to address the complex issues involved in metadata creation and maintenance. Applying traditional cataloging skills during descriptive metadata creation and enhancing data with applications for automated analysis and transformation—such as data mining, name and subject heading mapping, and batch processing—will improve the quality of the metadata in our repository and the efficiency with which it is created. The UH Libraries will continue to explore and experiment with new approaches to describing our digital objects and, with the metadata upgrade work outlined in this paper, we are laying the groundwork for the migration of our data to a more expansive semantic environment.

References

Art & Architecture Thesaurus. (2014). http://www.getty.edu/research/tools/vocabularies/aat/index.html/. Accessed July 26, 2014.

AutoHotkey. (2014). http://www.autohotkey.com/. Accessed May 29, 2014.

Handbook of Texas. (2014). http://www.tshaonline.org/handbook/. Accessed July 26, 2014.

Library of Congress Linked Data Service. (2014). http://id.loc.gov/. Accessed July 26, 2014.

Library of Congress Name Authority File. (2014). http://id.loc.gov/authorities/names.html/. Accessed July 26, 2014.

Library of Congress Subject Headings. (2014). http://id.loc.gov/authorities/subjects.html/. Accessed July 26, 2014.

OpenRefine. (2014). http://openrefine.org/. Accessed August 10, 2014.

Thesaurus for Graphic Materials. (2014). http://www.loc.gov/pictures/collection/tgm/. Accessed July 26, 2014.

Thesaurus for Use in College and University Archives. (2014). http://www.archivists.org/publications/epubs/thesaurus.asp/. Accessed July 26, 2014.

Thompson, Santi and Annie Wu. (2013). Metadata overhaul: upgrading metadata in the University of Houston Digital Library. Journal of Digital Media Management, 2(2): 137-147.

UHDL Metadata Dictionary. (2014). http://digital.lib.uh.edu/about/metadata/. Accessed August 7, 2014.

W3C. (2004). OWL Web Ontology Language Reference. Retrieved July 26, 2014 from http://www.w3.org/TR/owl-ref/#sameAs-def/.

Weidner, Andrew J. (2014). UHDL_Names_CDM. GitHub Repository. Retrieved May 29, 2014, from https://github.com/metaweidner/UHDL_Names_CDM/.

Weidner, Andrew J. (2014). UHDL_SubjectTopical_CDM. GitHub Repository. Retrieved May 29, 2014, from https://github.com/metaweidner/UHDL_SubjectTopical_CDM/.

Posters

Retaining Metadata in Remixed Cultural Heritage Objects

Jamie Wittenberg
University of Illinois at
Urbana-Champaign, USA
wttnbrg2@illinois.edu

Keywords: metadata; semantic web; remixing; linked open data

1. Context

Memory institutions have been working to incorporate features into their digital collections that empower users to take ownership of cultural narratives. The advent of technologies like annotation tools and crowdsourced tagging have allowed libraries, archives, and museums to promote user content as part of an institutional narrative, albeit a somewhat tertiary one (Salomon, 2013). Collecting institutions including the Smithsonian, MoMA, Australian Museum, and British Library have been developing initiatives that encourage users to remix openly available digital content. A remix appropriates components of existing resources and incorporates them into a new work.

This movement towards user-generated remixed content is cost effective for institutions and engaging for patrons. Increased interactivity is emblematic of the changing role of libraries, archives and museums (Reiskind 2012, pp. 6). The future of cultural memory institutions will be one that embraces collection diversity and incorporates user-generated material into institutional narratives. This is already happening in social media, crowdsourced tagging, API development, and remixing. Work to ensure that associated metadata is harvested along with media content is still in its naissance. Increased endorsement of remixing as a way of engaging with cultural heritage material requires a metadata infrastructure that can support description of remixed content in a way that is comprehensive, interoperable, and scalable.

2. Existing Standards

There are two primary obstacles preventing the development of such a model. The first is that even when comprehensive metadata is documented and available, current metadata standards do not describe content with sufficient specificity. Because remixes appropriate segments of items, rather than the entire item as a collection does, remixes require descriptions that are more granular. In order to accommodate the clipping and cropping nature of remixing, a more robust system of detailed object description is necessary.

The second obstacle is that metadata is often unidirectional. It is created for new items that may express relationships to existing records, but less commonly updated in existing records. To create metadata for remixes, metadata for original material would first need to be evaluated for its relevance to the new content. Metadata for each appropriated component part that makes up the remixed content should at minimum contain provenance, attribution, and descriptive information.

2.1. Descriptive Metadata Standards

In widely used descriptive metadata standards such as MODS and Dublin Core, relationships between items are FRBR-type hierarchical relationships. Remixes seem to occupy an unspecified space within the FRBR universe, because they appropriate and reuse items, rather than works or expressions. Remixes take a single physical instance of a manifestation and modify it. MODS and Dublin Core provide enough room in their structure that with some manipulation, it would be possible to approximate a description of a remix. This is especially true if the remix is an expression of the original work. However, some remixes might only incorporate minutiae of existing content, drawing it together to create an entirely new work. Neither the MODS RelatedItem attributes nor the Dublin Core Relation Type attributes express the relationship between source content and a remixed object that is a new work (LOC 2013; DCMI 2012). There is no possibility to include metadata touching on remixing actions, cardinality, or provenance. Given that this form of cultural production is not only becoming increasingly popular, but is being adopted into institutional narratives, there is a need for a metadata infrastructure that explicitly addresses remixed material (Fisher, 2007).

2.2. Event-Based Metadata Standards

Event based metadata standards such as CDWA, CIDOC-CRM and LIDO orient representation towards changes in the state of the item. These standards are better equipped than descriptive metadata schemas to manage the lifecycle data associated with cultural heritage material (Coburn 2010, pp.3-4). While event based standards offer the necessary process and provenance support to a remix metadata model, the scope of such standards is steered towards chain of custody-type changes such as the CIDOC-CRM Activity subclasses of Acquisition, Transfer of Custody, and Curation Activity (ICOM/CIDOC 2013 pp. 5). Remixed cultural heritage objects require a description that targets state changes in content production as well as lifecycle events after accession.

5. Future Work: Linked Data and Annotation Standards

Metadata for remixed objects must enable consistent description and attribution for all aspects of the work. Exploring Linked Open Data conceptualizations of aggregation and annotation such as the Open Annotation Data Model and the OAI-ORE Abstract Data Model offers insight into possibilities for structuring metadata associated with remixed cultural heritage objects (OAC 2013; OAI 2008). Such a structure must provide a descriptive framework for each component of a remix and would require an extensible model flexible enough that elements could be included from across domains. A standard that builds on Semantic Web concepts like the graph data model has the potential to provide that flexibility. This is an area that requires further research.

6. Conclusion

The profile of the heritage institution of the future is beginning to take shape, and it is characterized by ever-increasing interactivity, user customization, and widespread dissemination. Libraries, archives, and museums will be participatory, collaborative spaces with room for alternative narratives of heritage. Metadata structures and standards must adapt with these institutions. It is essential to the integrity of cultural heritage institutions that as traditional unilaterally created corpuses transition into inclusive and dynamic collections, descriptive infrastructures transition as well (Bertacchini, 2013, pp. 60). The movement towards enabling remixes of cultural heritage materials threatens existing metadata models because it requires systemic change in the granularity of descriptive metadata and in metadata creation workflows. The development of a metadata structure that can accommodate remixed content will help to ensure that libraries, archives and museums continue to fulfill their roles as stewards of cultural heritage content.

References

Bertacchini, E., & Morando, F. (2013). The Future of Museums in the Digital Age: New Models for Access to and Use of Digital Collections. *International Journal Of Arts Management*, 15(2), 60-72.

Coburn, E., Light, R., McKenna, G., Stein, R., Vitzthum, A. (2010). *LIDO - Lightweight Information Describing Objects Version 1.* Retrieved from http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

DCMI. (2012). *Dublin Core Metadata Element Set, version 1.1.* Retrieved from http://www.dublincore.org/documents/dces/.

Fisher, M. & Twiss-Garrity, B.A. (2007, March). Remixing Exhibits: Constructing Participatory Narratives With On-Line Tools To Augment Museum Experiences. *Museums and the Web 2007: Proceedings*. Toronto: Archives & Museum Informatics.

ICOM/CIDOC Documentation Standards Group. (2013). *Definition of the CIDOC Conceptual Reference Model*. Retrieved from http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf

Library of Congress. (2013). MODS Elelments and Attributes. *MODS User Guidelines ver. 3*. Retrieved from http://www.loc.gov/standards/mods/userguide/generalapp.html

Open Annotation Community Group. (2013). Open Annotation Core Data Model. Retrieved from http://www.openannotation.org/spec/core/20130208/index.html

Open Archives Initiative. (2008). ORE Specification - Abstract Data Model. Retrieved from http://www.openarchives.org/ore/1.0/datamodel

Reiskind, A. (2012). extraMUROS and the 21st century image library. *VRA Bulletin, 38*(2), 1. Retrieved from http://online.vraweb.org/vrab/vol38/iss2/4

Salomon, D. (2013). Moving on from Facebook. *College & Research Libraries News*, 74(8), 408-412.

Embedded Metadata—A Tool for Digital Excavation

Ana Cox
Phoenix Art Museum, USA
ana.cox@phxart.org

Keywords: embedded metadata; digital asset management; controlled vocabulary; collection management; data mapping

1. Introduction

In June of 2012, I commenced the weighty task of searching the far reaches of Phoenix Art Museum's digital storage spaces to import images into a recently acquired collection management system, The Museum System (TMS). Before my newly created position as Visual Resource Coordinator began, each department generated and stored assets with their own organizational system in digital silos. I excavated long forgotten folders on various servers and desktops, hunting for visual documentation of the art collection and past installations. Embedded metadata was used as a tool to identify subject matter of images and indicate which folders had been searched. These assets were then reorganized with a new file name convention and folder structure. This poster will discuss my method for using embedded metadata to track information about digital assets as well as challenges and opportunities for further development. This method could be implemented by other cultural organizations as a low cost approach to tracking basic metadata, content creators and copyright restrictions.

2. Implementation

The VRA XMP Info Panel, developed by the Visual Resource Association Embedded Metadata Working Group (VRA EMWG), was installed to view in Adobe Bridge and provides metadata fields that specifically pertain to cataloging art objects that the standard IPTC panel does not provide. The VRA Panel also adheres to controlled vocabularies such as Dublin Core and VRA Core. Thus, rules for cataloging were largely pre-established. The goal was to include only the most pertinent information for identifying the art object and how the file was created. The fields in Table 1 were identified to be the most useful.

TABLE. Metadata Fields.

Artwork	Image	Administration	Summary
Creator	Creator	Collection[1]	Description[2]
Title	Date	Cataloger	
Date	Source		
Medium	Copyright Restrictions		
Dimensions	Copyright Notice		
Repository	Custom Field: Image Type		
Description[3]	Custom Field: Document Type		
	Custom Field: Object Number		

[1] This field is used to assign curatorial area (controlled vocabulary).
[2] Caption information is concatenated from work fields.
[3] This field is used for additional notes about the artwork. For example, if multiple objects are included in the same image.

Only two custom fields were independently developed that were not included in the VRA Panel: Object Number (institutional object tracking number) and Image Type (controlled vocabulary: scanned transparency, reference image and professional collection photography).

3. Workflow

As digital silos were reviewed, labels in Adobe Bridge were utilized to mark which files and folders had been reviewed, which images were copied and cataloged according to the new file taxonomy and which images required subject identification. Where appropriate, object numbers were added as embedded metadata to assist with future identification. Once the files were copied to the new structure, the VRA Panel Export-Import Tool was used to transfer object metadata from TMS to Adobe Bridge via an Excel spreadsheet. Metadata was also embedded regarding how the image was created, suitability for publication and any copyright restrictions. As images were imported into TMS, this additional metadata was ingested into corresponding fields in the media record. Adobe Bridge provides several tools and features that allow the user to add embedded metadata to large batches of images as well as automated file renaming tools, which greatly improved the workflow.

4. Results

The hunt for these digital assets is ongoing, however after the initial survey spanning five months, I was able to import about 10,000 files into TMS, which is a 280% increase from the files imported into the previous collection management system, Argus. I also established procedures for cataloging and importing new assets. Today there are 16,708 media records in TMS. Overall, I have copied and cataloged approximately 74,000 files with embedded metadata into the digital archive. This number is growing daily.

5. Challenges and Opportunities

Using this method presented a few clear challenges and opportunities for development.

- Open source tools provided by the VRA EMWG make tracking digital assets through embedded metadata a low-cost, fast solution for digital asset management. For small to mid-size cultural organizations this method can effectively organize institutional history. However, in order to read every field in the VRA Panel, a staff member would need to download the panel and install it in the Adobe Creative Suite. If an organization does not already use Adobe products, there could be a cost barrier in acquiring this software and investing in staff training.
- The concatenated artwork caption appears in the description field in the standard IPTC panel, which can be read by a staff member using any tool that reads embedded metadata, such as Finder or Windows Explorer. This facilitates the ease of object identification; however, caption information is not static. For example, if the Registrar completes a vault inventory, is it worthwhile to correct all the updated measurements? Similarly, if the work of an artist moves into the public domain, is it worthwhile to update every image by this artist in the copyright restrictions field? Using this method for variable information could prove time consuming and requires constant attention and editing.
- The main benefit of using embedded metadata to track digital assets is that it is a tool to recognize images that have previously been ingested into a digital asset management system. For example, a staff member may create a copy of an image and rename the file to store in their digital silo. The embedded metadata is copied as well, thus providing a provenance for the file.
- Phoenix Art Museum does not currently use digital asset management software. If we were to move in this direction, the embedded metadata could easily be exported into an Excel spreadsheet and imported into a DAMS.

Despite these challenges, utilizing embedded metadata to track and describe digital assets is a low cost digital asset management solution for galleries libraries archives or museums. Embedded metadata is not only a useful tool for digital excavation, but can also provide opportunities as a starting point for a more nuanced digital asset management system.

Dublin Core to Ensure Interoperability between Models Generated by Tools of Species Distribution Modeling

Cleverton Ferreira Borba
University of Sao Paulo, Brazil
cleverton.borba@gmail.com

Pedro Luiz P. Correa
University of Sao Paulo, Brazil
pedro.correa@usp.br

Keywords: species distribution modeling; connection between tools; biodiversity informatics; ecological Informatics.

Abstract

This poster presents the use of the Dublin Core for tools that make species distribution modeling. As a case study, this poster proposes the use of the Dublin Core for there to be a connection between the models generated by tools of species distribution, contributing to the area for biodiversity informatics.

1. Introduction

The area of scientific research called Biodiversity Informatics is a new area of research that has received much attention in recent years because their results and innovations assist in decision making for conservation and preservation of biodiversity. According to Peterson et al (2010) this area is "challenged to meet the demand for support to biodiversity conservation technology".

The species distribution modeling makes it possible to verify the changes in species distribution, changes in populations and their diversity for a given period. However, studies show that currently modeling species distribution has become more complex (Soberón, Nakamura, 2009). And equally modeling tools require improvements to the application of new techniques and modeling strategies (Peterson et al. 2011).

One of the requirements is to ensure data interoperability between modeling tools. Interoperability means the ability of information exchange through a metadata standard. In this context the Dublin Core could help being adopted as a standard of data between models generated by the modeling tools distribution of species.

2. Dublin Core and their use in species distribution modeling

Currently the main link between the modeling of species distribution and Dublin Core is that modeling tools access different database platforms that use the Dublin Core standard for publishing and standardization of information.

The use of metadata standards such as Dublin Core also assists in collecting biodiversity data process because the data becomes public and through standardization is possible that the data is available on various platforms.

At this stage, display and availability of data collected, this poster proposes the use of the Dublin Core is further explored and used between models generated by modeling tools distribution of species.

3. Application of Dublin Core for connection between models of species distribution

The current structure of species distribution modeling tools is that they generate independent models and may not be used or reused by other tools. This makes the researcher / user having to use more than one tool to reach your goal.

The idea is that with the Dublin Core standard, other standards, and an ontology, it is possible to create a connection between the tools, ensuring interoperability between them, as we can see in Figure 1.

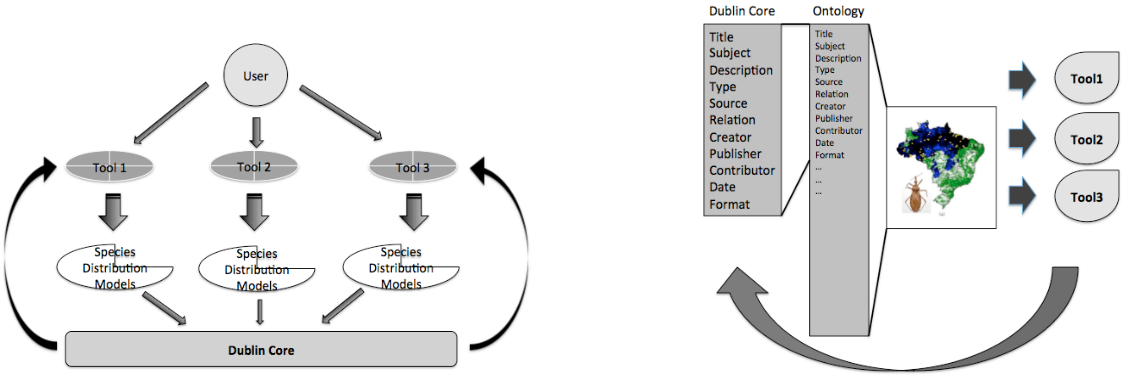


FIG. 1. Using Dublin Core for connection between modeling species distribution tools.

The proposed use of the Dublin Core standard for interoperability between models generated by tools of species distribution modeling is the use of the main elements of the Dublin Core. Every model generated must have a title, subject, description, type, source, relation, creator, publisher, contributor, date, format, etc.; this will ensure interoperability of basic information between the generated models. From this information an ontology with the main elements of the model should be a priority for the connection between modeling tools.

4. Conclusion

In conclusion of this part of the research it is possible to realize that the use of the Dublin Core can assist in the process to ensure interoperability between models generated by modeling tools distribution of species.

The Dublin Core standard has been one of the references regarding standardization for data availability and data visualization, and this would have a strong acceptance of the researchers for this standard is adopted as a party basis for a connection between modeling tools distribution species.

4.1. Future research

As future work, we suggest: creating an ontology based on the Dublin Core standard to ensure interoperability between tools; evaluation of the use of the Dublin Core in the tools and portals that help biodiversity conservation.

References

DCMI. (2012). Dublin Core Metadata Element Set, version 1.1: Reference description. Retrieved April 12, 2014, from http://dublincore.org/documents/dces/

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.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. & Araújo, M. B. (2011) Ecological Niches and Geographic Distributions. United Kingdon: Princeton University Press. 328.

Soberón J. & Nakamura M. (2009) Niches and distributional areas: Concepts, methods, and assumptions. PNAS 106(2).

Project Report: Building Bridges to the Future of a Distributed Network: From DiRT Categories to TaDiRAH, a Methods Taxonomy for Digital Humanities

Jody Perkins
Miami University
perkintj@miamioh.edu

Quinn Dombrowski
University of California
quinnd@berkeley.edu

Luise Borek
Technical University of Darmstadt
borek@linglit.tu-darmstadt.de

Christof Schöch
University of Würzburg
christof.schoech@uni-wuerzburg.de

Keywords: digital humanities; taxonomies; research methods; ontologies; RDF; Resource Description Framework; LOD; Linked Open Data

1. Project Background

This poster presentation traces the development and application of 'TaDiRAH' (Taxonomy of Digital Research Activities in the Humanities), a shared taxonomy of digital humanities research goals and methods (e.g. capture, enrichment, analysis), objects (e.g. data, images, manuscripts), and techniques (e.g. cluster analysis, encoding, topic modeling) created for the purpose of bridging the divide between related digital humanities hubs.

Earlier efforts to establish centralized hubs of information relevant to digital humanities (DH) have proven unsustainable over the long term. These comprehensive hubs (such as arts-humanities.net, a European initiative which previously aggregated information about events, jobs, news, projects and tools) are currently being re-designed with a smaller scope and more focused curation. However, this smaller scope comes with the risk of decontextualization—a digital humanities project is best understood through the intersection of its subject matter, methodologies and applications, not all of which are captured by any single site.

An example of a focused directory is the DiRT (Digital Research Tools) Directory, an established, well-regarded source of information about tools available to support scholarship in the humanities. DiRT is currently undergoing a new phase of development, with the goal of making information about digital tools available outside the DiRT directory itself using RDF and APIs.[1] However, the ad-hoc set of categories that have been used to organize tools on DiRT since its inception are of no utility outside DiRT itself. Adopting a shared taxonomy would provide a means to connect DiRT's tool data with related information provided by other sites.

2. Development Process

Early in 2013, as part of an effort to improve usability of the site, members of the DiRT Steering Committee/Curatorial Board conducted an analysis of DiRT's categories and free-form tags. Shortly thereafter we began a series of discussions with the DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities-Germany) team that was developing a taxonomy for their 'Doing Digital Humanities' Zotero bibliography. Recognizing our common goal, we formed a transatlantic collaboration around the task of developing a shared taxonomy.

In the process of developing TaDiRAH we drew from three primary sources: 1) the *arts-humanities.net taxonomy* for DH projects, tools, centers, and other resources, especially as it has

[1] http://dirtdirectory.org/development

been expanded by digital.humanities@oxford in the UK and DRAPIer (Digital Research and Projects in Ireland); 2) the *DiRT categories* for digital research tools, re-launched under Project Bamboo in the US but now continuing on after the end of that project; and 3) the *scheme used by the DARIAH 'Doing Digital Humanities' Zotero bibliography* to organize literature on all facets of DH. These resources were mapped, analyzed and distilled into their essential parts, producing a simplified taxonomy of two levels: eight top-level "goals" that are broadly based on the steps of the scholarly research process and a number of lower-level "methods" associated with each goal. In addition, there are two separate open ended lists of digital humanities research "objects" and "techniques" that can be freely associated with higher level methods.

In September 2013, and again in January 2014, we opened a draft version of the taxonomy for public comment and received a tremendous amount of feedback from the DH community. The response shows the ongoing relevance of a task that has been under discussion in digital humanities circles since John Unsworth introduced his concept of 'scholarly primitives' in 2000. We hope that one outcome of this presentation will be to extend the conversation beyond the boundaries of the DH community.

3. Challenges and Future Work

This presentation will also cover some of the challenges encountered during TaDiRAH's development, including: selection of terms that facilitate consistent application vs. terms that represent entities in a more precise manner[2], avoiding conflation of concepts, reconciling terms against existing taxonomies, minimizing redundancy, balancing theoretical "correctness" on one hand against the necessity of adopting commonly used terms to ensure findability on the other (e.g. visualization + geospatial coordinates object vs. "mapping"), and responding to thorough (and sometimes conflicting) feedback from the digital humanities community.

We will also present several use cases based on the shared taxonomy, demonstrating how it will work to serve both task and user-oriented endeavors. Applying TaDiRAH to actual directories will provide an opportunity to assess the degree to which it can accommodate real-world data. In the coming months we will conduct a comprehensive review of all DiRT tool entries, adding terms from the TaDiRAH taxonomy. DHCommons will also add TaDiRAH terms to project profiles based on existing free-form metadata. Information from DiRT and DHCommons will be exposed using RDF, making the content available as linked open data, as well as through APIs that are currently under development.

The "Doing Digital Humanities" bibliography curated by DARIAH-DE has already implemented the TaDiRAH taxonomy. The Zotero-based bibliography is using "collections" (similar to subfolders) for the seven broad goals, and the tags for the research activities, objects and techniques. Each entry is tagged with at least one activity and one object to enable a faceted browsing of the bibliography, starting with either research activities or objects. Most recently TaDiRAH has been adopted by two additional DARIAH initiatives: the Digital Humanities Course Registry and the Training Materials Collection (Schulungsmaterial-Sammlung).

DARIAH-EU has committed to using this taxonomy as a basis for their development of a more complex ontology of digital scholarly methods, and we are also engaged in ongoing dialog with other ontology initiatives, including NeDiMAH's (Network for Digital Methods in the Arts and Humanities) work around scholarly methods. Our goal is to share at least high-level categories with NeDiMAH's ontology, so that objects (projects, tools, articles, etc.) classified using our taxonomy can be automatically "mapped" to some level of the NeDiMAH ontology, and vice versa.

[2] While the use of specific terms supports precision, the use of more broadly defined terms tends to provide better support for consistent application, collocation and recall. In the context of search, precision and recall are often inversely related.

The projects and collections that adopt TaDiRAH will also inform its evolution. TaDiRAH can be found online at GitHub[3], where we will be using the issue tracker to collect further feedback to be incorporated into future revisions. A SKOS version soon to be available on the GitHub site and a SPARQL endpoint through a TemaTres instance are currently in development. We expect that TaDiRAH will continue to evolve as a relatively flexible scheme of associated scholarly methods, techniques and object types that can be applied to a variety of DH resources.

Acknowledgements

The authors would like to recognize the work of Matt Munson, who was a member of the TaDiRAH Coordinating Committee during its initial development phase.

References

Digital Humanities Course Registry, DARIAH-DE Cologne / CLARIAH-NL Rotterdam. (2014). Retrieved from http://dhcoursereg.hki.uni-koeln.de/

DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities-Germany). (2014). Retrieved from https://de.dariah.eu/

DHCommons. (2014). Retrieved from http://dhcommons.org/

DiRT (Digital Research Tools Directory). (2014). Retrieved from http://dirtdirectory.org/

Doing Digital Humanities - A DARIAH Bibliography. (2014). Retrieved from https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography/items/order/creator/sort/asc

NeDiMAH (Network for Digital Methods in the Arts and Humanities). (2014). Retrieved from http://www.nedimah.eu/

Schulungsmaterial-Sammlung, DARIAH-DE Würzburg/Cologne (2014). Retrieved from https://de.dariah.eu/schulungsmaterial-sammlung

TaDiRAH (Taxonomy for Digital Research Activities in the Humanities). (2014). Retrieved from https://github.com/dhtaxonomy/TaDiRAH and http://tadirah.dariah.eu.

Unsworth, John. (2000). Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? London: King's College London. Retrieved May 16, 2014 from http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

[3] http://github.com/dhtaxonomy/TaDiRAH

Metadata Workflows Across Research Domains: Challenges and Opportunities for Supporting the DFC Cyberinfrastructure

Adrian Ogletree
Drexel University
adrianogletree@gmail.com

Keywords: metadata workflows; metadata generation; DataNet Federation Consortium (DFC); research data; cyberinfrastructure.

1. Introduction

This poster presents research results from a survey studying metadata workflows. In the context of this study, a 'metadata workflow' is defined as a workflow that generates metadata for a data collection. The following research question guided this investigation: Where are people (and automated processes) creating metadata in the data life cycle, and what could be done to improve the quality?

2. Background

Metadata is necessary to find, use, and properly manage scientific data. Sharing metadata workflows across different communities is thus crucial for promoting data interoperability and reuse. The DataNet Federation Consortium (DFC) is a project within the NSF Office of Cyber-Infrastructure DataNet initiative. One widespread problem that the DFC seeks to address is the unfortunate reality that "many scientific fields lack a common integrated data infrastructure, which often results in non-standardized, local data management practices" (Akmon, 2011, p. 330-331). Carole Goble, Robert Stevens, Dave De Roure, and others have made significant contributions to the study of e-science workflows and reproducibility. In addition, Taverna and Kepler are two open-source, community-driven, scientific workflow management systems with large user bases in the eScience community (Taverna; Kepler). However, data management needs vary substantially across disciplines. Willis, Green, and White (2012) call for future research to examine in greater detail the "community-specific practices and workflows as well as constraints caused by the technological environment and trends at the time of scheme creation" (p. 1517).

3. Methodology

A survey was distributed via e-mail to the DFC listserv in order to better understand how scientific metadata is created. DFC scientists, researchers, and data curators involved in any aspect of creation or use of scientific metadata were invited to participate in this study.

4. Results and Discussion

Fourteen (14) participants responded to the survey, representing a 34% response rate (the DFC listserv contains 41 members). They were affiliated with eight different DFC project partners: the Ocean Observatories Initiative (OOI),[1] the iPlant Collaborative,[2] the Odum Institute for Research in Social Science,[3] the National Oceanic and Atmospheric Administration (NOAA),[4] the Renaissance Computing Institute (RENCI),[5] the University of Virginia, the Data Intensive Cyber

[1] http://oceanobservatories.org/

[2] http://www.iplantcollaborative.org/

[3] http://www.odum.unc.edu/odum/home2.jsp

[4] http://www.noaa.gov/

[5] http://www.renci.org/

Environments (DICE) Center,[6] and the School of Information and Library Science at the University of North Carolina at Chapel Hill. The participants’ fields of study included hydrology, biology, climatology, ecology, library sciences, computer science, engineering, social sciences, and information science. The composition of the participants’ positions were as follows: 2 professors, 1 associate professor, 1 assistant professor, 1 postdoc researcher, 1 doctoral student, 2 master’s students, 2 administrators, 1 software engineer, 1 scientific analyst, and 1 IT project team lead (one participant did not respond to this question). Five (5) of the participants had 5 to 10 years of research experience.

The following types of data were created or used in the participants’ research: observational data (7), papers (7), simulation data (4), laboratory experimental data (3), “other” (3), and field experimental data (1). Participants were asked to select all that apply. Observational data has the most long-term value for researchers because it is often unique, irreplaceable, or costly to collect (Anderson, 2004).

Figure 1 below shows metadata creation by a person and metadata creation or capture by a computer. Participants were asked to select all that apply; for instance, some researchers add metadata at every point within the data collection process. Eight (8) of the participants who responded to this question manually create metadata before data is collected, 10 manually create metadata during data collection, and all 12 manually create metadata afterward. Only 2 of the participants report that computer-generated metadata is created before data is collected; 9 report that automated metadata creation occurs during or after data collection, with one respondent selecting “other,” who had no automated metadata collection. Data management best practices recommend that data documentation happen at the very beginning of the research project, before data collection. However, these results indicate that more scientific metadata is created during or after the data collection process than before, and that few researchers take advantage of automated metadata generation workflows.

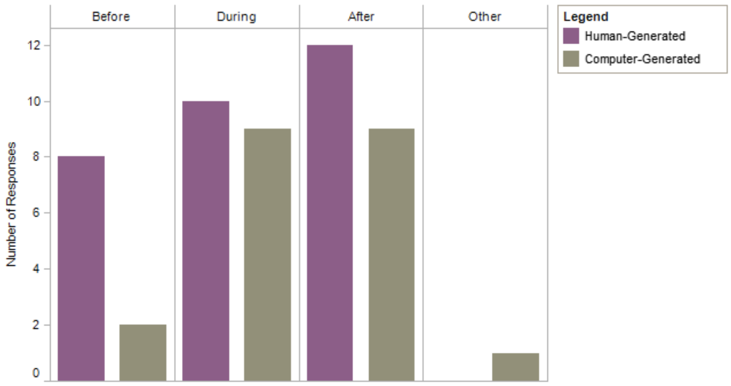


FIG. 1. Metadata creation by humans and automated processes.

Six (6) of the participants reported that their organization has a specified standard in place for creating metadata. The following metadata schemes were used: Dublin Core (7), “Other” (7), FGDC (2), NetCDF Climate and Forecast (CF) (2), “Don’t know” (1), EML (1), and “No standard scheme is used” (1). Participants were asked to select all that apply. Six (6) of the participants who selected “other” named the following metadata schemes: free tag AVU in irods, MIxS, DDI (2), WaterML, and GML. Based on the survey results, many different metadata schemes were used, consistent with Greenberg’s (2005) study of digital repositories that “hundreds of metadata schemes [are] being used, many of which are in their second, third, or *n*th iteration” (p. 18).

[6] http://dice.unc.edu/

When asked what information another researcher would need to reproduce their research, responses include: information about workflows, highly specialized knowledge, software, or equipment, and/or algorithms and parameters used. Similarly, Borgman (2012) observes that research reproducibility requires "the precise duplication of observations or experiments, exact replication of a software workflow, degree of effort necessary, and whether proprietary tools are required" (p. 17). Without contextual information and high-quality metadata, even "open" data is unusable.

5. Conclusions

Overall, the results met expectations based on other similar studies of scientists' data management practices and perceptions (Akers, 2013; Anderson, 2004; Borgman, 2012; Chavan & Penev, 2011; Greenberg, 2005). The following list represents the key findings of this survey:

- More than half (58%) of participants create or use observational data
- Metadata is more likely to be created after data collection
- Scientists and researchers suffer from a lack of awareness of metadata standards
- Data sharing is complicated by the need for highly specialized knowledge, software, and/or equipment in order to reproduce research

This study makes a contribution towards methods of survey design for the purposes of studying metadata workflows. Although the responses to this survey represent multiple scientific disciplines, positions, and institutions, this study was limited by the small sample size. Future research should include larger populations, and different research domains can be categorized in order to study the similarities and differences of data management needs between communities. Another area of interest for the DFC is the ability of the iRODS data grid to capture the provenance information associated with execution of a workflow. This research could be useful for creating a definition of a sufficient context to enable re-use of data.

Acknowledgements

I would like to thank Jane Greenberg and Reagan Moore for their guidance and support. The DFC DataNet is supported by the National Science Foundation, OCI-0940841.

References

Akers, Katherine G. and Jennifer Doty. (2013). Disciplinary differences in faculty research data management practices and perspectives. The International Journal of Digital Curation, 8(2), 5-26. doi:10.2218/ijdc.v8i2.263

Akmon, Dharma, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom. (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.

Anderson, William L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. Data Science Journal, 3, 191-201.

Borgman, Christine 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

Greenberg, Jane. (2005). Understanding metadata and metadata schemes. Cataloging & Classification Quarterly, 40(3-4), 17-36. doi:10.1300/J104v40n03_02

Kepler. Retrieved from https://kepler-project.org/

Taverna. (2014). Retrieved from http://www.taverna.org.uk/

Willis, Craig, Jane Greenberg, and Hollie White. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505-1520.

A Cooperative Project by Libraries and Museums of China: Metadata Standards for the Digital Preservation of Cultural Heritage

Ying Feng
CALIS Administrative Center,
China
fengy@calis.edu.cn

Long Xiao
Peking University Library,
China
lxiao@lib.pku.edu.cn

Keywords: cultural heritage; digital preservation; metadata standards; descriptive metadata; administrative metadata; preservation metadata.

This poster introduces a project that aims to build metadata standards for digital preservation of cultural heritage. Research and demonstration will be made by collaborative effort among seven libraries and museums.

1. Background and Objectives

In addition to preserving cultural heritage, the objective of digitization of cultural heritage is to share cultural heritage and related knowledge in an effective, rapid and convenient manner in context of networked environment, to provide information and knowledge services relating to the cultural heritage. At present, a number of museums in China have been digitalizing their culture heritage. However, it is difficult to integrate, share, and apply these digital outcomes due to the lack of uniform standards. At the same time, a large number of cultural heritage remain to be digitized. To avoid repeated problems, a standard metadata system for digital cultural heritage is required for comprehensive information organization, description, management and preservation. Additionally, other standards such as classification system for cultural heritage is also needed for building knowledge database of digital cultural heritage. Thus, it is urgent to establish a uniform metadata standards for digital preservation of cultural heritage.

Metadata Standards for Digital Preservation of Cultural Heritage is one of key research areas and sub-project of the Research and Demonstration Project on Standard Systems and Key Standards for Digital Preservation of Cultural Heritage which is funded by the Ministry of Science and Technology of China in this year. The objectives focus on the demands for business management, digitization, management of digital content, long-term preservation of digital content, and the establishment of a knowledge database for cultural heritage. The research will be based on existing metadata standards and use the application logic of the digital preservation of cultural heritage as its starting point. Then construct the metadata framework, core standards, description standards, administrative and preservation standards, and application technology specifications for digital preservation of cultural heritage, thereby standardizing metadata generation during digitization and preservation of cultural heritage, supporting and promoting the construction of digital preservation for cultural heritage, and driving the research, presentation, applying, and development of cultural heritage preservation.

Composition of the project team: Seven entities are involved in research as follows: Peking University, the Palace Museum, Dunhuang Research Academy, National Library of China, Zhejiang University, Tsinghua University and University of Science and Technology of China, with Peking University being the team leader.

Project development timeframe: 2014–2017.

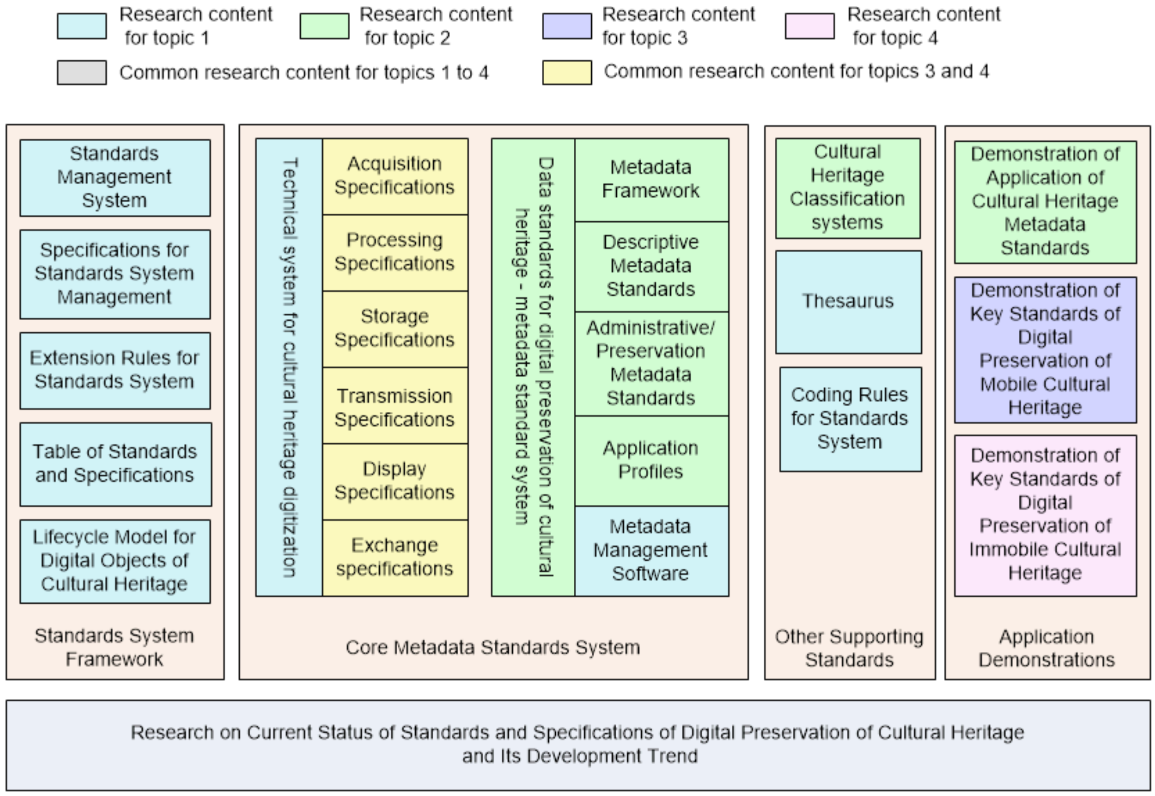

Research content for topic 1
Research content for topic 2
Research content for topic 3
Research content for topic 4
Common research content for topics 1 to 4
Common research content for topics 3 and 4
Standards Management System
Specifications for Standards System Management
Extension Rules for Standards System
Table of Standards and Specifications
Lifecycle Model for Digital Objects of Cultural Heritage
Standards System Framework
Technical system for cultural heritage digitization
Acquisition Specifications
Processing Specifications
Storage Specifications
Transmission Specifications
Display Specifications
Exchange specifications
Data standards for digital preservation of cultural heritage - metadata standard system
Metadata Framework
Descriptive Metadata Standards
Administrative/ Preservation Metadata Standards
Application Profiles
Metadata Management Software
Core Metadata Standards System
Cultural Heritage Classification systems
Thesaurus
Coding Rules for Standards System
Other Supporting Standards
Demonstration of Application of Cultural Heritage Metadata Standards
Demonstration of Key Standards of Digital Preservation of Mobile Cultural Heritage
Demonstration of Key Standards of Digital Preservation of Immobile Cultural Heritage
Application Demonstrations
Research on Current Status of Standards and Specifications of Digital Preservation of Cultural Heritage and Its Development Trend

FIG. 1. Composition of the Research and Demonstration Project on Standard Systems and Key Standards for Digital Preservation of Cultural Heritage

2. Key Barriers

The cultural heritage metadata standards under this research project must be able to describe the basic information of cultural heritage and meet the needs of business activity, while fulfilling the application requirements for digitizing cultural heritage and constructing the knowledge database. Flexibility, scalability and applicability also need to be considered. The key barriers and difficulties are as below:

1. The research and development of metadata framework for the digital preservation of cultural heritage. This is a fundamental technical issue for establishing the metadata framework for digital preservation of cultural heritage, which will directly affect the scientificity and rationality. The difficulties include:

- Revealing of properties, digitization, business activities related, knowledge database construction of mobile and immovable cultural heritage, as well as study and analysis on corresponding application requirements of cultural heritage metadata standards.
- Abstracting application requirements and building relationships among the concepts, as well as constructing a metadata information model that meets the requirements of the client.

2. The establishment of cultural heritage classification system. A cultural heritage classification system needs to account for the characteristics of both the digital objects of cultural heritage and physical entities. The scientific characteristics and practicality of each situation will directly influence the segmentation and design of the metadata standards description. Given the complex nature of cultural heritage, it is relatively difficult to construct a scientific and rational classification system for them.

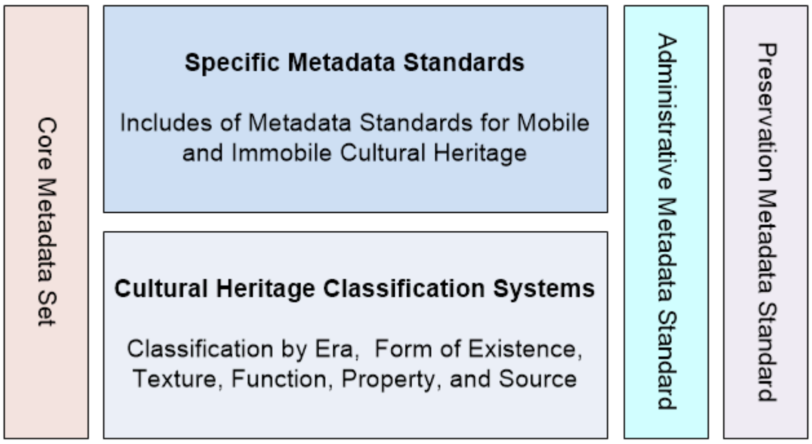

Core Metadata Set
Specific Metadata Standards
Includes of Metadata Standards for Mobile and Immobile Cultural Heritage
Cultural Heritage Classification Systems
Classification by Era, Form of Existence, Texture, Function, Property, and Source
Administrative Metadata Standard
Preservation Metadata Standard

FIG. 2. The relationship between the cultural heritage classification systems and metadata standard system.

3. Design of descriptive metadata system and specific metadata standards. The difficulties include the following:

- In order to meet different application requirements for cultural heritage metadata, modular, scalable, generic and customized descriptive metadata system and specific metadata standards are needed.
- How to make use of and integrate the various types of digital contents of cultural heritage already digitized to build foundation for implementing information sharing and the overall revealing of cultural heritage metadata.

4. Research and design of administrative and preservation metadata standards. Difficulties of abstracting and generalizing the application requirements of administrative and preservation metadata arise because of different business processes and management approaches among different cultural organizations. It is also difficult to design practical and scalable framework for administrative and preservation metadata.

5. The research and development of metadata application profiles which is shown in Figure 3.

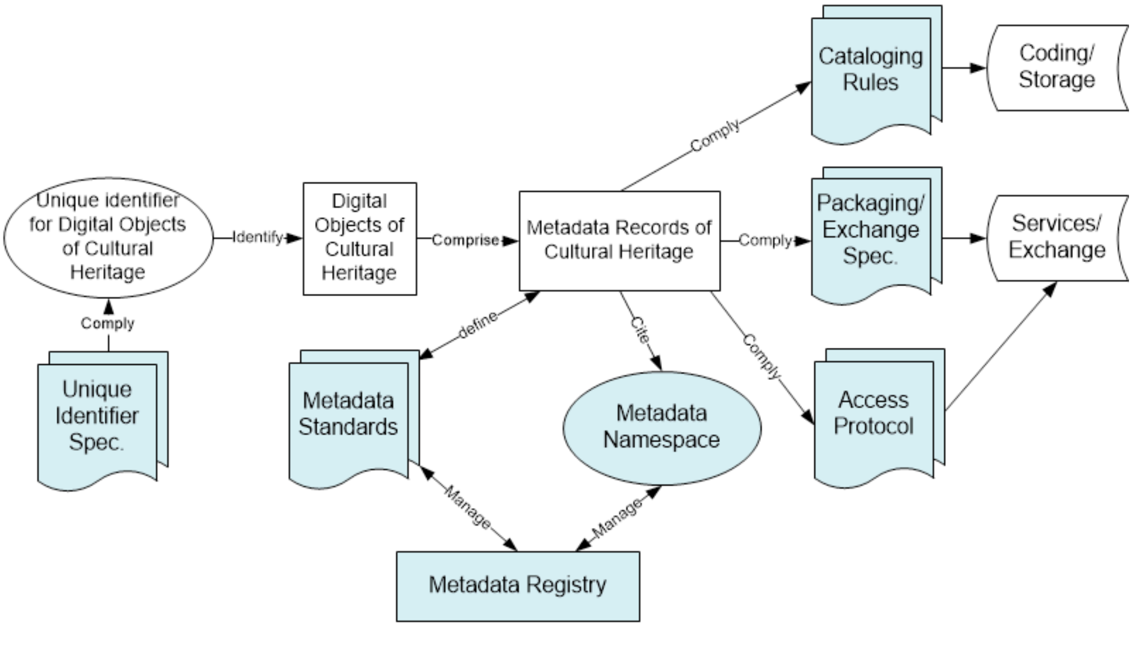


FIG. 3. Composition of metadata application profiles.

3. Design Principles and Expected Results

It focuses on the digital objects of cultural heritage in conjunction with the physical entities while designing cultural heritage metadata. Meantime, the following principles will also be considered, includes simplicity and accuracy, specificity and versatility, scalability and sustainability, interoperability and openness, user requirements and applicability.

The following are expected results.

- The metadata framework for the digital preservation of cultural heritage: includes general principles of cultural heritage metadata, metadata system for cultural heritage, metadata information framework for cultural heritage, core metadata set and its application guidelines, descriptive metadata application specification, and specific metadata design principles for cultural heritage.
- Classification systems for cultural heritage, for both digital and physical objects.
- Specific metadata standards for cultural heritage: includes 12 specific metadata standards and their cataloging rules as well as application guidelines for mobile cultural heritage, 7 specific metadata standards and their cataloging rules as well as application guidelines for immobile cultural heritage.
- Administrative and preservation metadata standards for the digital preservation of cultural heritage: includes metadata framework, element set and application guidelines for administrative metadata and preservation metadata for cultural heritage.
- Application profiles of metadata standards for the digital preservation of cultural heritage: includes metadata identification system, encoding rules, metadata packaging and exchange specifications, access protocol and open mechanisms.

References

CIDOC CRM (2013). Definition of the CIDOC Conceptual Reference Model. Retrieved June 10, 2013, from http://www.cidoc-crm.org.

MIDAS Heritage (2012). MIDAS Heritage: the UK Historic Environment Data Standard. Retrieved July 2, from http://www.english-heritage.org.uk/publications/midas-heritage/

Undressing Fashion Metadata:
Ryerson University Fashion Research Collection

Naomi Eichenlaub
Ryerson University,
Canada
neichenl@ryerson.ca

Marina Morgan
Ryerson University,
Canada
marina.morgan@ryerson.ca

Ingrid Masak-Mida
Ryerson University,
Canada
ingrid.masakmida@ryerson.ca

Keywords: fashion; metadata schema; metadata mapping; cataloguing; digital collections; digitization; Dublin Core; VRA Core.

1. Abstract

The purpose of this poster is to provide insight into the processes involved in making a unique fashion research and teaching collection discoverable in an online environment at Ryerson University. The online collection will provide a means for the users to identify what artifacts are available for research purposes and facilitate teaching in the classroom. The poster will highlight effective metadata standards and elements, cross-domain metadata uses, metadata mapping and implementation

2. Introduction

Ryerson University Fashion Research Collection project, a collaboration between the School of Fashion at Ryerson University and RULA (Ryerson University Library and Archives), consists of creating an online collection of images and metadata representing several thousand donated garments and accessories, designer clothing and millinery donated from private collections dating back to the latter part of the nineteenth century and early 20th century.

The key goals of this online collection are to promote research, teaching and learning at Ryerson University, and to connect with a broader community by building scholarly, online exhibitions. Once finalized, it will be used as a pedagogical tool and it will inspire fashion students and scholars to undertake research into fashion history.

3. Background

Ryerson University Library and Archives has partnered with the Ryerson School of Fashion to increase access to a unique collection of fashion items. The collection was housed in unfavourable conditions in a locked room in the library for many years and was relatively unknown to students. It was recently relocated to a series of rooms in the School of Fashion building. The collection is now in the process of being curated by its collection coordinator, who received a grant to digitize a portion of the collection in 2012. Currently, only very limited amount of information is available about the Ryerson Fashion Research Collection through a blog and a Pinterest site. Initially, a sample of the collection was loaded on to Pinterest as a means of both engaging students as well as exposing the collection to the world. The social media platform, however, has limited search functionality and virtually no descriptive metadata beyond an item description box.

Zeng (2009) asserts that the physical access restrictions common to most collections of historical fashion result from "delicate artifacts and by the inaccessible nature of many costume collection storage facilities". The online collection, however, will increase access from what was once multiple sets of excel spreadsheets of described items onto a searchable platform that will allow students, faculty and staff to have a more robust discovery, searching and browsing experience.

4. Research Significance

The research significance of the Ryerson Fashion Research Collection is three-fold. First, it will provide a venue for a greater expanse of fashion online exhibits, a pedagogical tool that will allow Ryerson students to learn and research, to foster students' interaction and participation, and to explore a rich yet previously inaccessible fashion collection at Ryerson University. Second, it will allow us to build on and implement future specific collections, to foster a connection between external and internal users, to promote and improve online access that would add value to the existing collection. Third, digital access will preserve the valuable collection, but at the same time will allow researchers, students, and the public to have "visual access to an entire collection without needlessly disturbing the garments and their accessories" (Zeng, 1999).

5. Metadata Implementation and Challenges

Very little has been written or published about the digitization of fashion collections and specifically about appropriate metadata schema for optimizing access and discovery of fashion object collections. The question of appropriate descriptive elements for use in fashion collection metadata records was noted by Marcia Lei Zeng in her article *Metadata elements for object description and representation: a case report from a digitized historical fashion collection project* specifically due to the three-dimensional nature of fashion artifacts (Zeng, 1999).

As Lampert and Chung (2011) argued, before developing and designing a digital collection, there are various technical questions to consider, such as thoroughly assessing various feature sets of different systems and making informed decisions to seek out appropriate solutions. Consequently, we evaluated and analyzed several web-publishing platforms, both proprietary and open source, and metadata standards that would better fit our criteria. Simplicity of installation of the software and metadata ingest were very important to us, particularly since we were working within a fairly short timeline. As well, metadata adaptability and interoperability, import and export functionality of specific standard data formats, flexible approaches to various plug-ins, (specifically the OAI-PMH Harvester) factored in to our selection criteria.

Judging against our expectations, we evaluated three possibilities that would best fit the selection criteria mentioned above: ICA-AtoM, SharedShelf (ARTstor), and Omeka.

	Metadata Standard	**Customizable**	**One to Many Relationship**	**OAI-PMH**	**Cost**
Atom	Various	No	No	Yes	Free
SharedShelf (ARTstor)	VRA Core	Yes	No	Yes	Subscription
Omeka	Dublin Core VRA	Yes	Yes	Yes	Free

Fig. 1 Comparison of possible web publishing platforms and metadata standards

AtoM (Access to Memory) is a web-based, dynamic open source application for standards-based archival description and access, allowing various import and export formats, and supporting ICA and non-ICA standards (RAD, Dublin Core, and MODS). Shared Shelf is a media management software that enables management, storage, use, and publishing institutional and faculty media collections within their institution, or publicly on the Web. Though highly customizable, complex and flexible, the platform did not allow for multiple image batch loading per individual item described (one-to-many relationships).

Omeka, a free, flexible, and open source web-publishing platform, on the other hand, allows the expansion of its core functionality with existing plugins to create maps, to allow users to tag favorites, and to create dynamic and robust online exhibits, thus tying closely with the pedagogical requirements of this project.

There are several other products on the market that offer features similar to Omeka. However, when it comes to providing the rich visual context or exhibiting collections, as today's web users would expect, these platforms may be less effective, often difficult to adopt, and more expensive to maintain than Omeka. Motivated to create digital collections due to the educational imperative to share their collections with the public, the academic world is often facing "restricted budgets and staffing issues" as Sauro (2009) argues. Faced with the same budget and staffing restrictions, we decided to go with the most flexible and cost effective solution for our project, Omeka. The decision was also based on the variety of features that Omeka offers. As highlighted by Kucsma, Reiss, and Sidman (2010), Omeka allows strong and flexible approach to metadata representation, straightforward plug-in deployment, customs creation of item types, and the addition of the full set of Dublin Core properties to the existing Dublin Core element set, including element refinements and supplemental elements.

Metadata element selection and metadata mapping was the next challenge. Selecting the metadata elements and refining the specifications is closely tied to the end-user usage patterns and item description choices. The stakeholders (users, faculty, and digital collection creators) have different ideas about what is useful in the collection. We learned that faculty use the collection for their own research as well as to enhance classroom learning; external researchers, visiting scholars, and curators are interested in a particular designer, period or type of artifact (e.g. 19th century hair accessories). We also learned that students seek access to the collection in different ways: to establish the specific material of a garment, or identify the type of stitching or other manufacture processes. Although the search strategy depends on the research question being asked, in general the primary search terms would be for a particular type of garment (corset, dress, coat, hat), period (1920s, 1950s, 1960s), designer (Balenciaga, Dior, Balmain), construction type (bias cut, inset sleeves), colour (yellow, orange, red), or textile (silk, linen, cotton). Consequently, we had consulted with the curator in order to determine the metadata elements highlighting the benefits of certain metadata elements. This possible usage pattern directly influenced the item description and metadata elements. What fits the faculty curricula vs. what fits the interest of the student or researcher directly impacted those choices.

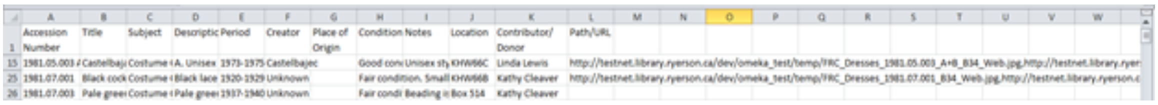
Zeng (1999) references difficulty of locating appropriate text to use as a title, an issue we faced as well. Discussion of the requirement for a title in each record was necessary, as it was understood that students would often be searching by accession number instead of title. Consequently, the curator of the Ryerson Fashion Collection created titles for each item using her expert knowledge in the field, adding mostly general terms such as "evening dress", followed in most cases by slightly more specific terms including the gender, colour, or shape of the garment, for example "Green wool men's tailcoat with black satin lapel and black wool vest".

Equally important when working with the metadata for these fashion items is the information needs of the students using the collection. In terms of providing subject access to the collection, after some discussion regarding the merits of additional subject access points we agreed that we would use the Art & Architecture Thesaurus (Getty Research Institute) (AAT). We had also considered using the Thesaurus of Graphic Materials (TGM) or Library of Congress (LC) Subjects, but our examination of fashion headings in the AAT revealed that it had better coverage in terms of fashion subject specificity. For example, the AAT has a term for dresses (garments) under which there are more than a dozen narrower terms including chemise dresses, coat dresses, gowns, jumpers (dresses), maxi dresses, midi dresses, muumuus, overdresses, etc. However, there are also limitations. For example, in using the AAT as it currently exists, certain commonly used terms for garments such as the word "pants" or "tunic" cannot be used. Consequently, to make items more discoverable, we used the tagging option in Omeka.

There were a number of other metadata fields that posed some challenges in order to meet the information needs of the students, while respecting the descriptive standards we were following. Rich description and details ended up mostly being mapped to one DC element: DESCRIPTION. VRA Core elements such as MATERIAL, MEASUREMENTS, CULTURE, OR STYLE/PERIOD would seem more appropriate given the specificity of this collection. Omeka does have the option of implementing the VraCoreElementSet plugin developed by the Scholars' Lab at the University of Virginia Library. Given the time and staffing constraints we haven't been able to configure and test it, however, this is something that we could develop in the future.

Batch image loading and one-to-many relationships was another challenge we faced. To be able to accomplish this task, we had two options: one was to use the OAI data to create a CSV instead of using it to import directly into Omeka. This would allow us to add a column to the CSV file with the location of the files for each item. The second option was to loop through the images and identify the correct object ID, i.e. the description to which the image

will be linked. A script looped through each line of the CSV file, stored the accession number into a variable, and for each accession number it looped through the filenames, adding only the matching filenames at the end of the line in the CSV. When opening the resulting CSV, a new column containing the matching filenames was created, thus allowing a smooth batch loading of images and item description, including the metadata.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Accession Number	Title	Subject	Descriptic	Period	Creator	Place of Origin	Condition	Notes	Location	Contributor/ Donor	Path/URL											
15	1981.05.003 A	Castelbaja	Costume (	A. Unisex	1973-1975	Castelbajac		Good cond	Unisex sty	KHW66C	Linda Lewis	http://testnet.library.ryerson.ca/dev/omeka_test/temp/FRC_Dresses_1981.05.003_A+B_B34_Web.jpg,http://testnet.library.ryer											
25	1981.07.001	Black cock	Costume (	Black lace	1920-1929	Unknown		Fair condition. Small		KHW66B	Kathy Cleaver	http://testnet.library.ryerson.ca/dev/omeka_test/temp/FRC_Dresses_1981.07.001_B34_Web.jpg,http://testnet.library.ryerson.c											
26	1981.07.003	Pale gree	Costume (	Pale gree	1937-1940	Unknown		Fair condi	Beading is	Box 514	Kathy Cleaver												

FIG. 2 Example of extra column added after executing the scripts

The last challenge was the collection's discoverability. Necessary for a quality user experience, the Fashion Research Collection should be seamlessly available via the library's discovery layer. However, this collection is not yet integrated into the library's discovery environment. In order for this collection to be discovered, a record in the library catalogue and one in the University's repository will be created.

6. Future Research

Looking to the future, a number of international digital fashion collection projects provide inspiration for the possibility of a similar Canadian initiative. The Europeana Fashion Portal is a three-year project to aggregate best of Europe's fashion collections and has a number of goals including improving interoperability and developing a specialized Fashion Thesaurus.[1] Australia also has a national initiative called the Australian Dress Register which showcases pre-1975 dress with Australian provenance and encourages museums and private collectors to "research their garments and share the stories and photographs while the information is still available and within living memory".[2]

Of particular relevance to future directions for online fashion resources is the possibility of incorporating interactive functionality and social features into collections to allow for user-generated content (Lampert, 2011). Allowing users to contribute their knowledge about historical items of little-known provenance, for example through tagging, can be an effective way to gather information that collection curators might otherwise miss. Moreover, incorporating some of the functionality of social curation sites such as Pinterest that allow for users to create their own personal digital collections is another possible future direction. The Omeka software has a

[1] http://blog.europeanafashion.eu/about/).

[2] http://www.australiandressregister.org/about/

number of plugins that offer a level of user interaction, such as the Comments, Exhibit Builder and MyOmeka, the latter allowing for item favouriting.

7. Conclusions

The Fashion Research Collection is a study collection consisting of several thousand artifacts including garments, accessories, and ephemera including photographs, magazines, and patterns. This collection is intended to support the research activities of the students and faculty at Ryerson University as well as means of engaging the outside community. This project met its overarching goals of increasing access and discoverability to a unique collection of mixed-provenance but mostly Canadian fashion items. Furthermore, the collaboration between the School of Fashion and Ryerson University Library and Archives allowed for subject matter experts in fashion, cataloguing and metadata standards to collaborate on a project that will provide community members and the public-alike with access to a tool for research, teaching, and learning.

References

Artstor Digital Library. (2014). Shared Shelf Features. Retrieved from

http://www.artstor.org/shared-shelf/s-html/features.shtml

Dublin Core Metadata Initiative. (2012). DCMI Metadata Terms. Retrieved from

http://dublincore.org/documents/dcmi-terms/

ICA-AtoM. Retrieved from https://www.ica-atom.org

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

Lampert, C. & Chung, S.K. (2011). Strategic planning for sustaining user-generated content in digital collections. Journal of Library Innovation, 2(2), 74-93.

Omeka Plugins. (2014). Retrieved from http://omeka.org/add-ons/plugins/

Sauro, C. (2009). Digitized historic costume collections: Inspiring the future while preserving the past. Journal of the American Society for Information Science and Technology, 60(9), 1939-1941. doi:10.1002/asi.21137

Tzoc, E., & Millard, J. (2011). Technical skills for new digital librarians. Library Hi Tech News, 28(8), 11-15. doi:10.1108/07419051111187851

Valentino, M. L. (2010). Integrating metadata creation into catalog workflow. Cataloging & Classification Quarterly, 48(6-7), 541-550. doi:10.1080/01639374.2010.496304

VRA Core Schemas and Documentation. (2007). Retrieved from http://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf

Zeng, M. L. (1999). Metadata elements for object description and representation: A case report from a digitized historical fashion collection project. Journal of the American Society for Information Science, 50(13), 1193.

Best Practice Posters & Demonstrations

Best Practice Poster: MARC to schema.org: Providing Better Access to UIUC Library Holdings Data

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

Michael Norman
University of Illinois at Urbana-Champaign, United States
manorman@illinois.edu

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

William Weathers
University of Illinois at Urbana-Champaign, United States
weathrs2@illinois.edu

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

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

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

Keywords: MARC, schema.org; MARCXML; bibliographic description; MODS; holdings data.

1. Introduction

Taking advantage of the Web as a means for disseminating large datasets, libraries have begun publishing their bibliographic metadata on the Web—e.g., the University of Michigan,[1] the University of Florida,[2] and Harvard University.[3] Initially, most libraries focused on releasing their catalogs as MARCXML, however, MARC consists primarily of string data with few, if any, URIs linking to ontologies or related resources. MARCXML was not designed for use with RDF. Libraries are now experimenting with disseminating catalogs as linked open data in other serializations, e.g., OCLC,[4] and the British Library.[5] Semantics compatible with RDF are being used, but specific schemes vary. Detail about holdings associated with bibliographic descriptions is still lacking, e.g., the volumes of a described serial title held by the library are not enumerated. This last seems a significant omission given that libraries are uniquely positioned to provide this information. The University of Illinois at Urbana-Champaign (UIUC) Library has released 5.5 million bibliographic catalog records that include detailed local holdings information to allow consumers to know exactly which volumes or parts of the creative work described are available at UIUC. MARCXML serializations are available for downloading now. MODS serializations enriched with links to name and subject authorities and RDF serializations (using schema.org semantics) will soon be available. This poster reports on the development of workflows for this project, on the multiple formats of catalog metadata being made available through these workflows, and on the lessons learned to date.

[1] http://www.lib.umich.edu/library-information-technology/open-access-bibliographic-records-available-download-and-use

[2] http://www.uflib.ufl.edu/catmet/creativecommons.html

[3] http://openmetadata.lib.harvard.edu/bibdata

[4] http://www.worldcat.org/

[5] http://bnb.data.bl.uk/

2. MARCXML with physical holdings information

As a first step, we created MARCXML bibliographic descriptions for each physical volume the library holds with selected volume-specific information (e.g., barcode) recorded in the 955 local data field. With a simple VB.NET program, we collapsed volume-level records associated with a single bibliographic entity into one bibliographic record that contains all holding and item level information for associated volumes and parts in repeated MARC 852 data fields as shown in Figure 1.

```
<marc:datafield tag="852" ind1="0" ind2=" ">
 <marc:subfield code="a">IU</marc:subfield>
 <marc:subfield code="b">Rare Book &amp; Manuscript Library [non-
circulating]</marc:subfield>
 <marc:subfield code="h">099</marc:subfield> <!-- classification number -->
 <marc:subfield code="i">Ab3</marc:subfield> <!-- cutter -->
 <marc:subfield code="p">30112066264109</marc:subfield> <!-- barcode -->
 <marc:subfield code="t">1</marc:subfield> <!-- copy number -->
</marc:datafield>
```

FIG 1: Example of MARC XML 852 data field used to record physical holdings

3. MODS Transformation & Adding Links

The transformation of MARCXML with holdings information in 852 data fields into MODS is based on the Library of Congress (LC) MARC to MODS recommendations.[6] (We differ slightly from the LC mapping recommendations in how we treat enumeration/chronology, copy number, and barcode.) Each 852 data field is mapped to a MODS <location> element. 852 subfield a is mapped to <location> sub-element <physicalLocation>; all other 852 subfields map to sub-elements of a single <copyInformation> element, within the <holdingSimple> subelement of <location>. Figure 2 displays the 852 data field of Figure 1 transformed to MODS.

```
<mods:location>
  <physicalLocation displayLabel=”Institution Code”>IU</physicalLocation>
  <holdingSimple>
    <copyInformation>
      <subLocation> Rare Book &amp; Manuscript Library [non-circulating]</subLocation>
      <shelfLocator>099 Ab3</shelfLocator>
      <note displayLabel=”Copy Number”>1</note>
      <note displayLabel=”Barcode”>30112066264109</note>
    </copyInformation>
  </holdingSimple>
</mods:location>
```

FIG 2: MARC 852 data field transformed to MODS

After transforming MARCXML records to MODS, a Python script is invoked to search VIAF for URIs matching values in the MODS <name> element, as transformed from MARCXML data fields 100, 110, 111, 700, 710, 711, and 720. When found, URIs are added to the MODS <name> element replacing the string values. When searching VIAF, we use complete name information, birth date, and death date (as available). Only exact matches in VIAF are recorded. The same script searches LCSH Linked Data Services[7] to find subject heading URIs, which are then also

[6] http://www.loc.gov/standards/mods/userguide/location.html

[7] http://id.loc.gov/

added to the MODS <subject> element. If no match is found, the text string remains as the value for the field.

4. Transformation to RDF and schema.org

The MODS metadata enriched with links to name and subject authorities are transformed into schema.org semantics. These are disseminated one-by-one as RDFa (within HTML styled for presentation to end-users), via bulk downloading (as RDF/XML or JSON-LD), and via a SPARQL endpoint. Transformation of bibliographic metadata from MODS to schema.org is straightforward (though arguably the distinction between work and manifestation is further blurred). However, transforming holdings to schema.org is challenging. Based on earlier experimentation at OCLC and our interpretation of relevant W3C Schema Bib Extend Community Group guidelines,[8] we mapped each holding as a schema.org <offer> entity.

Conclusion

The goal of this poster is two-fold:

- sharing with the community practices and workflow implementations developed at UIUC for disseminating traditional library data in multiple formats and serializations; and,
- gaining feedback on the mapping and modeling decisions made in transforming detailed MARC bibliographic and holdings data into linked open data.

[8] http://www.w3.org/community/schemabibex/wiki/Holdings_via_Offer

Best Practice Poster: The TR32DB Metadata Schema: A Multi-level Metadata Schema for an Interdisciplinary Project Database

Constanze Curdt
University of Cologne,
Germany
c.curdt@uni-koeln.de

Dirk Hoffmeister
University of Cologne,
Germany
dirk.hoffmeister@uni-koeln.de

Keywords: metadata; Dublin Core, research data; data repository; interdisciplinary

Abstract

The multi-level TR32DB Metadata Schema (Curdt, 2014) was designed and implemented with the purpose to describe all heterogeneous data, which are created by project participants of an interdisciplinary research project, with accurate, interoperable metadata. The metadata schema considers the interoperability to recent metadata standards and schemas. It is applied in the CRC/TR32 project database (TR32DB, www.tr32db.de), a research data management system, to improve the documentation, searchability and re-use of the data. The TR32DB is established for a multidisciplinary, long-term research project, the Collaborative Research Centre/Transregio 32 'Patterns in Soil-Vegetation-Atmosphere Systems: Monitoring, Modelling, and Data Assimilation' (CRC/TR32, www.tr32.de), funded by the German Research Foundation.

A key issue of research data management systems is the documentation of all research data with accurate metadata (Greenberg et al., 2013). This is particularly important for long-term research projects (Michener, 2006) and should follow recent metadata standards and schemas (Jensen et al., 2011). Consequently, the TR32DB Metadata Schema is designed in a multi-level approach combining several metadata schemas and standards, as well as data type and project specific metadata elements to describe all heterogeneous data. Metadata elements of Dublin Core are applied as a base schema. To meet the requirements of different TR32DB data types (data, geodata, report, picture, presentation, publication), the Dublin Core metadata elements are extended with further elements of metadata standards and schemes like ISO19115 Metadata Standard[1], INSPIRE[2], as well as elements of the Bibliographic Ontology[3] or the Event Ontology[4]. In addition, metadata elements of the DataCite Metadata Schema Version 2.2[5] are complemented. Furthermore, the TR32DB schema is expanded with own metadata properties corresponding to the TR32DB data types (e.g. measurement instrument and parameter), as well as to the CRC/TR32 background (e.g. specific keywords, themes). The schema specifies a defined number of metadata properties, including a core set of mandatory properties, as well as optional and automatically generated properties (e.g. metadata creator and date). In addition, available and TR32DB-specific controlled vocabulary lists are supported. A mapping to the applied metadata standards is provided for interoperability.

In detail, the TR32DB Metadata Schema is arranged in two layers: a general layer and a specific layer. The general layer enables the description of all data with basic details (e.g. title, description, creator, subjects). They are required for all data types. The specific layer complements the documentation of the data with specific metadata properties for each data type. For example, datasets from the TR32DB data type 'data' can be described with specific

[1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:326:0012:0030:EN:PDF, http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf

[3] http://bibliontology.com/

[4] http://purl.org/NET/c4dm/event.owl

[5] http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

properties, such as the temporal extent (e.g. start/end data), the lineage, the used measurement instrument (e.g. equipment group/method, model, manufacturer) and corresponding measurement parameter. Furthermore, the datasets from the TR32DB data type 'geodata' can be described with specific attributes, such as a temporal extent (e.g. start/end data), a lineage, the applied reference system or spatial resolution. In addition, datasets from the TR32DB data type 'report' can be described with additional attributes, like a report date, the report type (e.g. PhD report, master thesis, fellow report), the city or institution, where the report was created. Moreover, datasets from the TR32DB data type 'picture' can be described with a recorded date (e.g. start/end date), the name of the recording place, the recording method and details about the recording event (e.g. event type, name, location, website). Finally, the TR32DB data type 'publication' makes an exception, because different publication types require various attributes. Consequently, an 'article' can be described, for example, with a type of article (e.g. journal, magazine), publication source, publisher, volume, issue, pages, and page range. In contrast, an 'event paper' specifies information about the event, where the paper was presented. This includes the event name, the location, and period. In addition, details about the proceedings title, the editor, as well as the page range of the paper can be specified.

CRC/TR32 participants provide their metadata of a dataset by the TR32DB web-interface. A user-friendly, self-designed metadata input-wizard enables the entry of the metadata. The data search through metadata is available for all visitors of the TR32DB website by predefined, advanced, and map search functions. As a result, a detailed overview of all available metadata of a selected dataset is provided, which is arranged according to the TR32DB Metadata Schema.

Overall, the interoperable TR32DB Metadata Schema allows the accurate description of all heterogeneous data, generated by the CRC/TR32 participants. The multi-level approach enables a simple enhancement of the schema according to changing requirements of the project participants.

Acknowledgements

We would like to thank all colleagues involved in the design and implementation of the TR32DB. In addition, we gratefully acknowledge financial support by the CRC/TR32 'Patterns in Soil-Vegetation-Atmosphere Systems: Monitoring, Modelling, and Data Assimilation' funded by the German Research Foundation (DFG).

References

Curdt, Constanze. (2014). TR32DB Metadata Schema for the Description of Research Data in the TR32DB. Cologne, Germany: Transregional Collaborative Research Centre 32, Project Section Z1/INF, Institute of Geography, University of Cologne. Retrieved July 1, 2014, from http://dx.doi.org/10.5880/TR32DB.10.

Greenberg, Jane, Swauger, Shea, & Feinstein, Elena. (2013). Metadata Capital in a Data Repository. Paper presented at the International Conference on Dublin Core and Metadata Applications, DC-2013, September 2-6, 2013. Retrieved July 1, 2014, from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/189/86.

Jensen, Uwe, Katsanidou, Alexia, & Zenk-Möltgen, Wolfgang. (2011). Metadaten und Standards. In S. Büttner, H.-C. Hobohm & L. Müller (Eds.), Handbuch Forschungsdatenmanagement (pp. 83-100). Bad Honnef, Germany: Bock u. Herchen.

Michener, William K. (2006). Meta-information concepts for ecological data management. Ecological Informatics, 1, 3-7.

Best Practice Poster: Development of the EDDA Study Design Terminology to Enhance Retrieval of Clinical and Bibliographic Records in Dispersed Repositories

Ashleigh Faith
School of Library and Information Sciences
University of Pittsburgh, United States
anp114@pitt.edu

Eugene Tseytlin
Department of Biomedical Informatics
University of Pittsburgh School of Medicine, United States
eugene.tseytlin@gmail.com

Tanja Bekhuis
Department of Biomedical Informatics
University of Pittsburgh School of Medicine, United States
tcb24@pitt.edu

Keywords: clinical records; bibliographic records; dispersed repositories; medical terminaology: design process; study designs.

1. Background

Medical terminology varies across disciplines and reflects linguistic differences in communities of clinicians, researchers, and indexers. Inconsistency of terms for the same concepts and lack of machine-readable metadata impede discovery of information artifacts, such as records of clinical reports and scientific articles that reside in various repositories. To facilitate discovery, retrieval, and data sharing, the medical community maintains an assortment of terminologies, thesauri, and ontologies. Valuable resources include the US National Library of Medicine Medical Subject Headings (MeSH), Elsevier Life Science thesaurus (Emtree), and the National Cancer Institute Thesaurus (NCIT). It is increasingly important to identify medical investigations by their design features, as these have implications for evidence regarding research questions.

2. Purpose

Recently, Bekhuis et al (2013) found that coverage of study designs was poor in MeSH and Emtree. Based on this work, the EDDA Group at the University of Pittsburgh is developing a terminology of study designs. In addition to randomized controlled trials, it covers observational or uncontrolled designs.

3. Methods

Among the resources analyzed thus far, inconsistent entry points, semantic labels, synonyms, and definitions are common. The EDDA Study Design Terminology is freely available in the NCBO BioPortal (http://purl.bioontology.org/ontology/EDDA). Some of the preferred terms have several variants, definitions sometimes compete, as well as other concept identifiers useful for researchers. The beta version was developed using the Protégé ontology editor v.4.3 (http://protege.stanford.edu) and distributed as a Web Ontology Language (OWL) file. Dublin Core Metadata Initiative (DCMI) protocols are in place for recording overall terminology metadata and OWL annotations.

4. Results

At this preliminary stage, the term matrix consists of 171 class axioms consisting of study design terms, related terms, and publication types. When possible, class axioms were annotated with definition(s), incompatibility status, legacy term(s), controlled vocabulary resource unique identification, semantic type, and variant term annotation properties. In addition, revision metadata was also captured with editor annotations consisting of the team member who modified the class axiom and the date of modification. The following process was used for axiom enhancement (Figure 1):

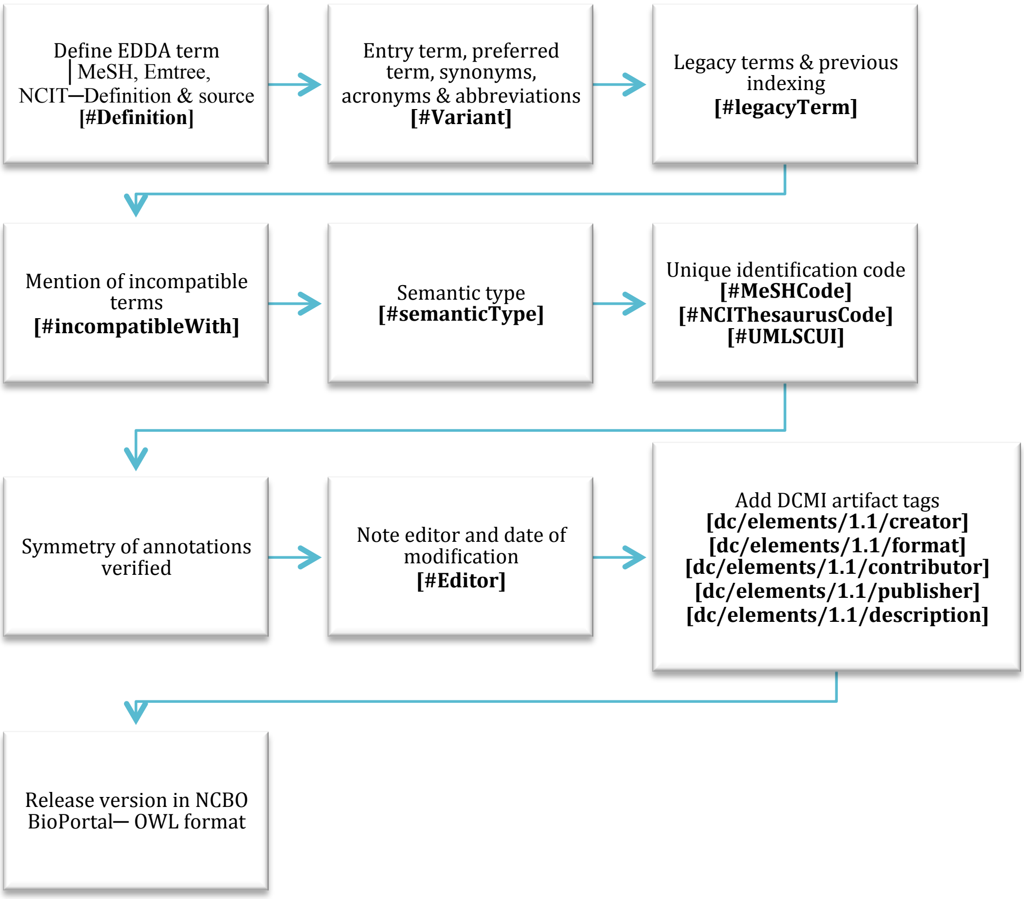

Define EDDA term │MeSH, Emtree, NCIT—Definition & source **[#Definition]**
Entry term, preferred term, synonyms, acronyms & abbreviations **[#Variant]**
Legacy terms & previous indexing **[#legacyTerm]**
Mention of incompatible terms **[#incompatibleWith]**
Semantic type **[#semanticType]**
Unique identification code **[#MeSHCode] [#NCIThesaurusCode] [#UMLSCUI]**
Symmetry of annotations verified
Note editor and date of modification **[#Editor]**
Add DCMI artifact tags **[dc/elements/1.1/creator] [dc/elements/1.1/format] [dc/elements/1.1/contributor] [dc/elements/1.1/publisher] [dc/elements/1.1/description]**
Release version in NCBO BioPortal— OWL format

FIG. 1. Design term annotation process.

Through the annotation process, a total of 2,381 axiom annotations were recorded. This included 51 MeSH, 33 NCIT, and 27 Emtree exact match access points to EDDA Study Design terms. Both MeSH and NCIT access points enabled information to be recorded. However, 12 of 27 Emtree access points did not result in any information because of insufficient information. Because NCIT cross-references other controlled vocabularies, 33 Unified Medical Language System (UMLS) resources also contributed to axiom annotations. A total of 123 definitions, 14 instances of incompatibility, 33 legacy terms, 95 semantic types, and 1,349 term variations were recorded.

5. Conclusions & Future Work

Identifying and retrieving reports of medical investigations by design features is increasingly possible, primarily through linking metadata. Further development entails adding definitions from other sources, mapping relationships among terms, and integrating terms from existing vocabularies, particularly the Information Artifact Ontology. A primary goal is to improve identification and retrieval of electronic records describing studies in dispersed data warehouses or electronic repositories.

Acknowledgements

This research was partially supported by the US National Library of Medicine (NLM), National Institutes of Health, grant number 5R00LM010943. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NLM.

Best Practice Poster: Normalizing Decentralized Metadata Practices Using Business Process Improvement Methodology: A Data-Informed Approach to Identifying Institutional Core Metadata

Emily Porter
Emory University, USA
eporter@emory.edu

Keywords: metadata standards; element sets; benchmarking; best practices; business process improvement; thematic analysis; quantitative analysis; qualitative analysis.

Environment, Context, and Techniques

The Emory University Libraries and Emory Center for Digital Scholarship have developed numerous digital collections over the past decade. Accompanying metadata originates via multiple business units, authoring tools and schemas, and is delivered to varied destination platforms. Seeking a more uniform metadata strategy, the Libraries' Metadata Working Group initiated a project in 2014 to define a set of core, schema-agnostic metadata elements relevant to local content types.

Quantitative and qualitative techniques commonly used in the field of Business Process Improvement were utilized to mitigate complex organizational factors. A key research deliverable emerged from benchmarking: a structured comparison of over 30 element sets, recording for each standard its descriptive element names, their required-ness, and general semantic concepts.

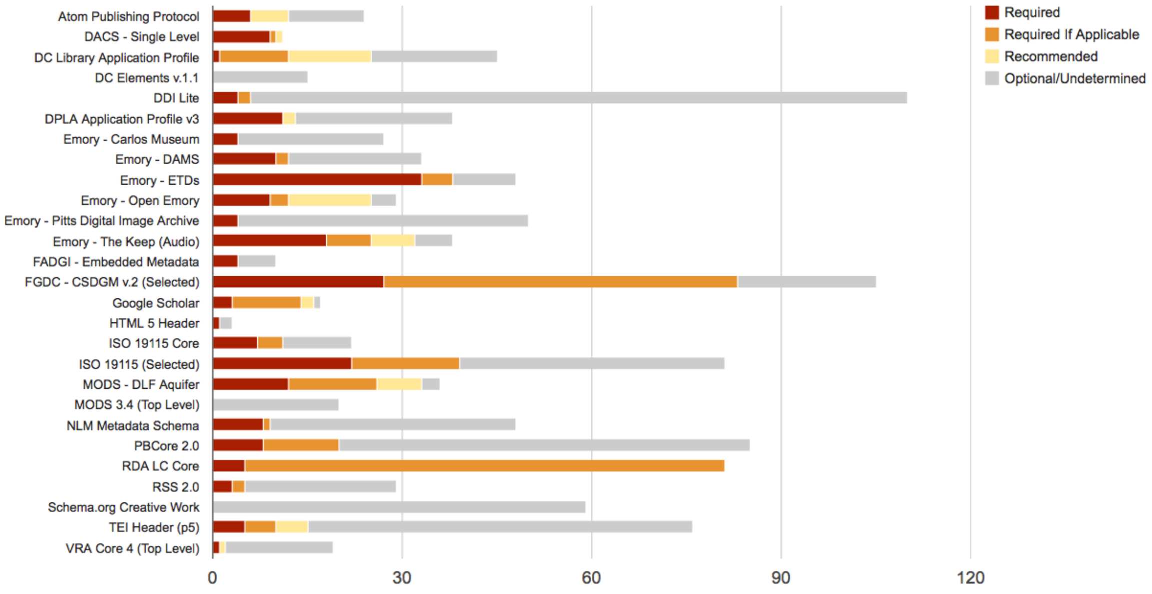


FIG. 1. Descriptive Elements by Schema/Standard: Quantity and Requirements (Selected Sources).

Additional structured data collection methodologies included a diagnostic task activity, in which participants with varying expertise created (simple) Dublin Core records for selected digital content. A survey of stakeholders provided greater context for local practices. Multiple public-facing discovery system interfaces were inventoried to log search, browse, filter, and sort options, and available web analytics were reviewed for user activity patterns correlating to these options.

Thematic analysis was performed on all benchmarking, system profiles, and web analytics data to map the results to a common set of conceptual themes, facilitating quantification and analysis. A weighted scoring model enabled the ranking of elements' themes: the highest scoring concepts then explicated as an initial set of core elements, mapped to relevant standards and schemas.

Acknowledgements

The Emory University Libraries' Metadata Working Group lent subject matter expertise to data collection and analysis (Brian Croxall, Jen Doty, Jason LeMay, Rebecca Koeser, Melanie Kowalski, Elizabeth Roke). Lars Meyer (Woodruff Library, Emory University) sponsored the project and has provided ongoing support. Emory's Office of Business Practice Improvement (Bill Dracos, Chris Rapalje, Jamie Smith, Ashley Teal) shared valuable techniques and expertise.

References

Atom Syndication Format – Introduction. (2007). Retrieved March 31, 2014, from http://atomenabled.org/developers/syndication/.

Berkman Center for Internet & Society at Harvard Law School. (2003). RSS 2.0 Specification (RSS 2.0 at Harvard Law). Retrieved March 31, 2014, from http://cyber.law.harvard.edu/rss/rss.html

Data Documentation Initiative. (2014). DDI Lite (Recommended Elements). Retrieved Month DD, YYYY, from http://www.ddialliance.org/sites/default/files/ddi-lite.html.

DCMI-Libraries Working Group. (2004). DC-Library Application Profile (DC-Lib). Retrieved Feb 13, 2014, from http://dublincore.org/documents/library-application-profile/.

Digital Library Federation. (2009). Digital Library Federation/Aquifer Guidelines for Shareable MODS Records, Version 1.1. Retrieved April 30, 2014, from https://wiki.dlib.indiana.edu/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf.

Digital Public Library of America. (2013). Metadata Application Profile, Version 3. Retrieved Month DD, YYYY, from http://dp.la/info/developers/map/.

Dublin Core Metadata Initiative. (2012). Dublin Core Metadata Element Set, version 1.1. Retrieved March 25, 2014, from http://www.dublincore.org/documents/dces/.

Embedded Metadata Working Group, Smithsonian Institution. (2010). Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images. Retrieved April 7, 2014, from http://www.digitizationguidelines.gov/guidelines/GuidelinesEmbeddedMetadata.pdf.

Federal Geographic Data Committee. (1998). CSDGM Graphical Representation. Retrieved April 23, 2014, from http://www.fgdc.gov/csdgmgraphical/index.html.

Google Scholar. (n.d.). Inclusion Guidelines for Webmasters. Retrieved April 17, 2014, from http://scholar.google.com/intl/en-US/scholar/inclusion.html#indexing.

International Standardization Organization (ISO). (2003). *Geographic information — Metadata*. New York: American National Standards Institute.

Library of Congress. (2013). MODS User Guidelines (Version 3). Retrieved February 28, 2014, from http://www.loc.gov/standards/mods/userguide/index.html.

Library of Congress. (2014). LC RDA Core Elements. Retrieved April 8, 2014 from http://www.loc.gov/aba/rda/pdf/core_elements.pdf

Miller, S. (2011). *Metadata for Digital Collections*. New York: Neal-Schuman Publishers, Inc.

PBCore. (2011). Elements. Retrieved April 21, 2014, from http://www.pbcore.org/elements/.

Schema.org. (n.d.). CreativeWork. Retrieved April 9, 2014, from http://schema.org/CreativeWork.

Society of American Archivists. (2013). Describing Archives: A Content Standard. Retrieved May 2, 2014, from http://files.archivists.org/pubs/DACS2E-2013.pdf.

Text Encoding Initiative. (2014). 2 - The TEI Header. Retrieved April 7, 2014, from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html.

U.S. National Library of Medicine. (2004). NLM Metadata Schema. Retrieved April 17, 2014, from http://www.nlm.nih.gov/tsd/cataloging/metafilenew.html.

VRA Core Oversight Committee. (2007). VRA Core 4.0 Element Description. Retrieved March 31, 2014, from http://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf.

W3C. (2014). 4.2 Document metadata. Retrieved July 1, 2014, from http://www.w3.org/TR/html5/document-metadata.html

Best Practice Poster: The NDL Great East Japan Earthquake Archive: Features of Metadata Schema

Akiko Hashizume
National Diet Library, Japan
hasizume@ndl.go.jp

Julie Fukuyama
National Diet Library, Japan
ju-fukuy@ndl.go.jp

Keywords: The Great East Japan Earthquake; archive; dcndl; ndlkn

1. Background

The Great East Japan Earthquake, which struck Japan on March 11, 2011, caused extensive damage in several parts of Japan and has affected Japanese society, culture and economy. Since immediately after the earthquake, the importance of passing on this historical experience to future generations has been pointed out in Japan and overseas. The Japanese government announced its basic policy towards the recovery from the earthquake. This policy pointed out the need to develop a system to collect, preserve and provide access to records of and lessons learned from the earthquake, tsunami and nuclear disaster.

Based on this policy, the National Diet Library (NDL), in conjunction with numerous other organizations throughout Japan, has developed the Great East Japan Earthquake Archive Project for the collection, preservation, and provision of information related to the earthquake.

2. The NDL Great East Japan Earthquake Archive

A portal site for this project was developed by the NDL and opened to the public on March 2013. Features available at the portal site include integrated searches of resources and reports on the earthquake and subsequent disasters produced by public institutions, private organizations, and mass media companies as well as research publications by universities, academic societies, and research institutions. The portal site has been named HINAGIKU, which means daisy in English.[1] This name is intended to convey an image of hope for the future and mutual concern in support recovery from the earthquake.

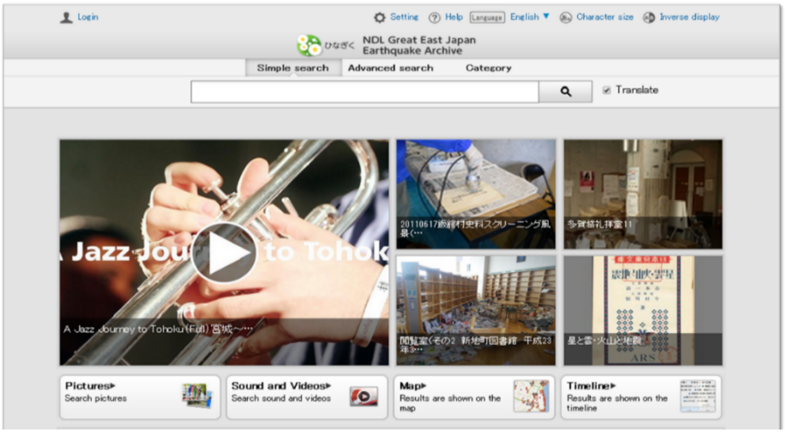


FIG. 1. Top page of the Great East Japan Earthquake Archive (HINAGIKU) (English version).

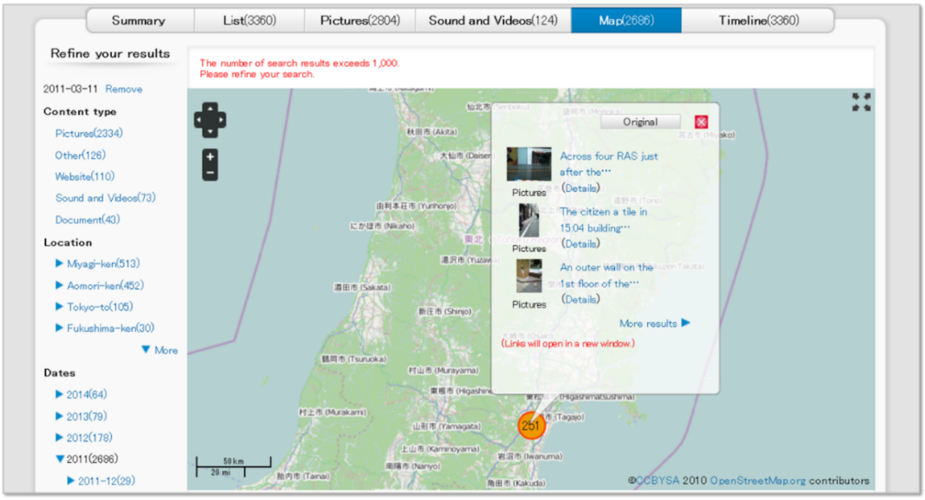
[1] HINAGIKU is an acronym of Hybrid Infrastructure for National Archive of the Great East Japan Earthquake and Innovative Knowledge Utilization.

HINAGIKU allows you to search the following resources. By the end of April 2014, the number of searchable records in HINAGIKU had reached 2,642,788.

TABLE 1: Resources collected in HINAGIKU.

Subject	Records of the Great East Japan Earthquake and the damage it caused, records of the affected areas before the earthquake, records of the restoration and reconstruction after the earthquake
	Records of aid activities by the national government, local municipalities, and other public organizations as well as records of aid activities by volunteer groups, non-profit organizations, and other private initiatives.
	Records of disaster prevention planning and academic research before and after the Earthquake as well as records of disaster prevention planning for the future
	Records of nuclear hazards resulting from the earthquake
	Records of earthquakes, tsunami, and other natural disasters from the past
	Records of the impact of past earthquakes on politics, economics, and society in Japan and around the world
	Records of the Great East Japan Earthquake and the damage it caused, records of the affected areas before the earthquake, and records of restoration and reconstruction after the earthquake
Format	Books, journals, newspapers, and other publications and digitized data
	Reports, research papers, news
	Websites of public and private organizations
	Images
	Video
	Audio (interviews, etc.)
	Fact sheets (observed data, geodetic data, etc.)

The user-friendly HINAGIKU interface includes a map display and a timeline display. Users interested in searching documents, images, video, and other digital material from a particular region can browse via the map display. Users interested in searching digital material chronologically search via the timeline. The time base can be changed to facility tracking the passage of time and reviewing the progress of reconstruction initiatives.



The number of search results exceeds 1,000.
Please refine your search.

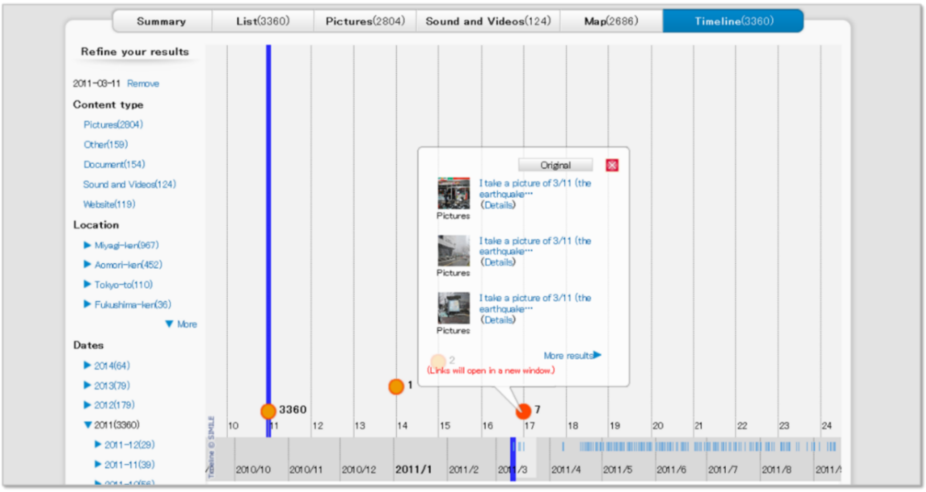
Across four RAS just after the… (Details)

The citizen a tile in 15:04 building… (Details)

An outer wall on the 1st floor of the… (Details)

(Links will open in a new window.)

FIG. 2. Map page of HINAGIKU (English version).



I take a picture of 3/11 (the earthquake… (Details)

I take a picture of 3/11 (the earthquake… (Details)

I take a picture of 3/11 (the earthquake… (Details)

(Links will open in a new window.)

FIG. 3. Timeline page of HINAGIKU (English version).

To enable integrated searches, HINAGIKU collects three types of metadata:

1. metadata on digital materials stored in HINAGIKU
2. metadata from the NDL's other databases, including the online catalog (NDL-OPAC)
3. metadata collected from other databases created by other organizations,[2] including those of local municipalities, universities, and mass media

To handle this metadata in HINAGIKU, we developed the Great East Japan Earthquake Archive Metadata Schema (NDLKN).[3] This schema is based on the National Diet Library Dublin Core Metadata Description (DC-NDL), which is our own metadata schema, based on the DCMES and DCMI Metadata Terms, for facilitating interoperation of metadata between libraries and

[2] The Examples of the cooperating organization with HINAGIKU is as follows:
CiNii Article by National Institute of Informatics
JAEA OPAC by Japan Atomic Energy Agency's Library
Digital Archive of Japan's 2011 Disasters by Edwin O. Reischauer Institute of Japanese Studies at Harvard University
East Japan Earthquake Picture Project by Yahoo!JAPAN etc.

[3] "NDLKN" is from 'NDL Knowledge infrastructure system metadata schema'.

related institutions in Japan. DC-NDL comprises NDL Metadata Terms as well as Application Profile and RDF Schema for NDL Metadata Terms.

Mechanical searches and harvesting of metadata are supported on HINAGIKU through Web API with SRU, OpenSearch, and OAI-PMH. API/SRU returns search results in RDF/XML.

3. The Great East Japan Earthquake Archive Metadata Schema (NDLKN)

NDLKN was created as an extension of DC-NDL, so that HINAGIKU could search the metadata not only of other institutions but also of NDL search systems, such as a discovery tool "NDL Search", which implement DC-NDL for metadata schema.

NDLKN comprises

1. 87 terms described in DC-NDL (dcndl:),
2. 33 terms described by W3C and adopted internationally (exif: etc.), and
3. 5 terms described originally in NDLKN (ndlkn:).

There were two major issues to solve in development of NDLKN. The first was coordination of metadata in various systems over multiple domains. The second was to satisfy requirements for archiving disaster records. NDLKN was developed to be a solution to these issues.

3.1. Coordinate with metadata of variable systems over domains

It was not possible to create metadata in the new NDLKN schema for existing domestic and foreign disaster record archive systems, because they held metadata in original schema. Therefore, we decided to harvest and keep metadata in the original schema in one storage and map this data to the NDLKN schema for storage for searching. We ask newly building archives to adopt NDLKN and to extend the terms according to the needs of each institution.

As mentioned above, NDLKN was extended from DC-NDL. The main differences between these schema are changes of the classes from [dcndl:Item] to [ndlkn:Resource] and from [dcndl:BibAdminResource] to [ndlkn:MetaResource]. The NDL Search, which implements DC-NDL based on FRBR model, holds terms for individual items in the class [dcndl:Item]. However, we felt that it would be difficult for organizations other than libraries to understand the concept of FRBR item, especially since HINAGIKU was intended to utilize digital materials such as images and videos more than books and journals found in traditional libraries. Also, we set [ndlkn:Resource] and changed the class [dcndl:BibAdminResource] to [ndlkn:MetaResource].

We also decided to store the URI of metadata providers in [dcterms:creator] of [ndlkn:MetaResource] class and the URI of the NDL in [dcterms:publisher]. We did this because we consider metadata providers to be primarily responsible for the metadata, which the NDL accepts and makes available.

We also assumed that the number of cooperating archives would continue to increase, and therefore it would be preferable to use identifiers for HINAGIKU metadata that would not require adjustment or reduction and would never be exhausted or overlap. As a result of these considerations, we adopted the UUID (Universally Unique Identifier)-RFC4122 and decided to add UUID to one file as a minimum unit.[4]

It is necessary to specify a license or terms of use for each resource that will be reused. Therefore, we decided to use [dcterms:license] for the information of the license and to adopt [cc:attributionURL] from Creative Commons Rights Expression Language to describe the name of the rights holders. Both of these are used in the form of URI.

Ex. 1: Creative Commons license

```
<dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/us/"/>
```

[4] The UUID Version 4 is a string of random 32 hexadecimal digits, so it is impossible to overlap the identifiers.

Ex. 2: Yahoo! JAPAN East Japan Earthquake Picture Project

<dcterms:license rdf:resource="http://archive.shinsai.yahoo.co.jp/contents/guide/"/>

Ex. 3: The NDL

<cc:attributionURL rdf:resource="http://www.ndl.go.jp/"/>

3.2. Meeting the needs of archiving disaster records

HINAGIKU functions not only as a web portal that enables integrated search for either digital or analogue resources but also as an archive that stores and preserves resources themselves with metadata. HINAGIKU archives digital materials such as images and videos at the moment. We considered the terms of NDLKN for each material types of objective resources.

NDLKN adopted [premis:formatName] and [premis:formatVersion] from PREMIS as terms for preservation technology. For images and videos recorded on digital cameras, we selected from Ontology for Media Resources by W3C, for example, concerning the recording location, [ma:createdIn] for the URI, [ma:locationLatitude] for latitude and [ma:locationLongitude] for longitude, and regarding the sound and video, [ma:samplingRate] for sound, [ma:frameRate] for video and [ma:duration] for playing time. We adopted the terms minimum amount necessary for images only [exif:width] for the width and [exif:height] for the height of the image from Exif data description vocabulary.

It is important for post-disaster surveys that resources such as images and videos have geospatial information. For this reason, we set terms not only for describing address, longitude, or latitude but also for distinguishing the objective space from the recording location of the resource. As for recording location, we adopted [v:street-address] and [v:postal-code] from Ontology for vCard. To describe the objective space of the resource, we described the value structure using [dcterms:spatial] and adopted [rdfs:label] for the name of the objective space, [v:region] for the prefecture, [v:locality] for the city, town and village, [v:street-address] for the street address, [v:postal-code] for the postal code, additionally [geo:lat] for the latitude, [geo:long] for the longitude from the terms of the Basic Geo (WGS84 lat/long) Vocabulary.

The temporal information is also important for disaster records. Therefore we described the date the image or video was recorded in [dcterms:created] and the date it was started to collect from a website in [dcndl:dateCaptured]. We recommend that values be stored in W3CDTF format, specifying by [rdf:datatype]. Furthermore, in HINAGIKU, metadata is mapped to W3CDTF format uniformly if possible, even if the provided metadata is not in W3CDTF format.

At the beginning of the development of the NDLKN, we assumed that it would be necessary to group the data by region, kind of disaster, or other characteristic useful to searching the data and displaying the search results. For this, we discussed to use the terms collection and item to represent a parent/child relationships in resources. However, after consideration, it became clear that it is almost impossible to describe collection uniquely. Therefore, we described both collection and item by [ndlkn:Resource] and chose to represent parent/child relationships in resources by connecting them with [dcterms:isPartOf] or [dcterms:hasPart].

HINAGIKU was initially intended to be an archive of the Great East Japan Earthquake. The target of the collection, however, includes records of earthquakes, tsunamis, and other past disasters, too, and other new archives might also be developed for future disasters. Based on these assumptions, we described [dcterms:coverage] to store the name and URI of disasters in order to describe the objective disaster of the resource.

4. Characteristic utilization examples of NDLKN in HINAGIKU system

We introduce several utilization examples of implementation of NDLKN in HINAGIKU system.

As HINAGIKU coordinates with domestic and foreign archive systems of disaster records, we assume that it would be necessary to confirm metadata schema definitions of its acquired time if cooperative organizations change their schema in the future. Therefore, HINAGIKU system stores its own URI in [dcterms:conformsTo] of [ndlkn:Resource] class and information of original metadata schema of providers in [ndlkn:sourceConformsTo] as internal term.

We also utilized the NDLKN terms [geo:lat], [geo:long], [ma:locationLatitude], [ma:locationLongitude] for the longitude and latitude of resources such as images and videos. HINAGIKU stores the latitude and longitude data automatically from either the name of the objective space or the recording location through the Yahoo! Geocoder API when the provided metadata does not have the value of latitude or longitude.

Web sites of the local governments of stricken areas and the Japanese government are also important as disaster records. The NDL has archived web sites for a long time by the WARP system and we have started to archive disaster related web sites with higher frequency after the Great East Japan Earthquake. As for the web sites, the titles (for example 'Sendai city') are not changed even if the content changes. Therefore it is necessary for searching and distinguishing the search results to add temporal information such as year, month, and date collected to the collected web sites. For this reason, HINAGIKU stores not only the value of title but also related information in [dcterms:title] in regard to the web sites collected by the WARP system. More specifically, we described to store the date started to collect too in [dcndl:dateCaptured] with [ ] after the title.

References

Basic Geo (WGS84 lat/long) Vocabulary. Retrieved April 28, 2014, from http://www.w3.org/2003/01/geo/wgs84_pos#.

Creative Commons. Creative Commons Rights Expression Language. Retrieved April 28, 2014, from http://creativecommons.org/ns#.

IFLA (2009). Functional Requirements for Bibliographic Records: FRBR. Retrieved April 28, 2014, from http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

NDL. (2011). National Diet Library Dublin Core Metadata Description (DC-NDL), version Dec. 2011. Retrieved April 28, 2014, from http://www.ndl.go.jp/en/aboutus/standards/index.html. The DC-NDL is DCMI Metadata Terms 2010-10-11 based metadata schema and compliant with the recommendations of the DCMI.

NDL Search. Retrieved April 28, 2014, from http://iss.ndl.go.jp/?locale=en.

PREMIS (2011). PREMIS Data Dictionary for Preservation Metadata, version 0.99. Retrieved April 28, 2014, from http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl.

The Great East Japan Earthquake Archive (HINAGIKU). Retrieved April 28, 2014, from http://kn.ndl.go.jp/.

W3C (2003). Exif data description vocabulary. Retrieved April 28, 2014, from http://www.w3.org/2003/12/exif/ns.

W3C. (2006). Ontology for vCard. Retrieved April 28, 2014, from http://www.w3.org/2006/vcard/ns.

W3C (2012). Ontology for Media Resources 1.0. Retrieved April 28, 2014, from http://www.w3.org/ns/ma-ont.

WARP. Retrieved April 28, 2014, from http://warp.da.ndl.go.jp/info/WARP_en.html.

Yahoo! Geocoder API (Japanese only). Retrieved April 28, 2014, from http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html.

Best Practice Poster: Reusing Legacy Metadata for Digital Projects: The Colorado Coal Project Collection

Michael Dulock
University of Colorado
Boulder, USA
michael.dulock@colorado.edu

Keywords: metadata; legacy metadata; digital libraries; digital collections; archives; Dublin Core; McBee cards; keyslot cards; edge-notched cards; metadata repurposing; hidden collections

1. Introduction

Libraries and other cultural institutions are increasingly focused on efforts to unearth hidden and unique collections. Yet the metadata describing these collections, when such exist, may not be in an immediately useable format. In some cases the metadata records may be as exceptional as the materials themselves. This poster describes research underway into how libraries can repurpose metadata in archaic formats using the Colorado Coal Project Collection[1] slides as a case study.

Metadata in outdated formats, whether analog or digital, are a mixed blessing for metadata practitioners when creating digital collections. On the one hand, practitioners are happy to have pre-existing descriptive information to accompany their materials, eliminating the need to re-describe a collection or the items it contains. On the other hand, a lot of work may be required to convert that metadata into a form that can be used by their digital systems. Examples of legacy metadata include archival finding aids in typescript, catalog cards, handwritten inventories, out-of-date database software, and other more exotic formats. The metadata thus preserved can provide a wealth of information for users of a digital collection, but first the data must be moved from its old format into a newer, digital system. Various tools, such as database conversion software or OCR (optical character recognition) applications, can be used to convert metadata. But those tools are not fool-proof. A text captured using OCR may still require manual quality checking, since OCR software may not be able to correctly interpret the inconsistencies of typescript. Even metadata captured in a spreadsheet may not be immediately useable. Manual intervention is required to separate different values in cells that contain multiple data points, for instance.

2. Background

The Colorado Coal Project Collection documents the history of coal mining in the western United States, primarily focusing on Colorado in the early 20th century. The original project was conducted between 1974 and 1984 by Eric Margolis and Ron McMahan of the Institute of Behavioral Science at the University of Colorado Boulder. The two researchers documented the history, technology, and lives of coal miners in Colorado through photographs and interviews with miners, community members, and historians to discuss topics ranging from mining camp life and immigration to working conditions, labor unions, and strikes. The physical collection, housed at the University of Colorado Boulder Archives, comprises over one hundred video and audio files of interviews, scores of transcripts, and over four thousand slides depicting mining life.

The slides are accompanied by over four thousand McBee cards, a manual computing format that saw occasional use for recordkeeping in the mid-20th century (McCoy, 1965; Rabinow, 1958;

[1] http://libcudl.colorado.edu:8180/luna/servlet/UCBOULDERCB1~76~76

Smith & Schnall, 1980). These cards contain written notes as well as punches around the edge which indicate various features of the slides such as locations, dates, and technical details. Transferring this rich metadata from thousands of cards into a workable digital format was a challenge. The poster examines the process of transferring the metadata recorded on these arcane cards to a 21st century digital library collection, utilizing a combination of student labor, Metadata Services staff, MS Excel, and careful quality control.

3. Methodology

The first part of the metadata transfer process was capturing the metadata on the cards in an electronic format that could then be manipulated. The data was recorded in a consistent manner according to a classification key included with the cards. Each card was divided into sections: text in the interior of the card recorded the slide number, title, date, and description information, image quality, and restriction/rights notes; a series of numerically-coded holes (locations for punches) were arranged around the edge of the card. These, too, were divided into sections according to type: decades, structures, historical notes, "general" notes; states and regions; "general categories"; and technical notes. (See poster for a card and key images.) Each numbered pinhole was assigned a value on the key. Categories on the cards were mapped to metadata elements from the Dublin Core Metadata Element Set (DCMI, 2012). The Metadata Librarian built an Excel spreadsheet to capture the card metadata by category, which could then be crosswalked to Dublin Core (DC). The spreadsheet had one column for each category (slide number, decades, etc.), with each row representing a single card. Multiple data points would be entered in a single column but separated by a delimiter so the Metadata Librarian could later create one column for each entry. A key was added at the top of the spreadsheet indicating valid values for each category (text, 1-14, L0-L8, etc.). The spreadsheet would be filled out with data exactly as it appeared on the card, including numeric codes.

The Metadata Services Department hired three student workers to manually transfer the data. Each would be expected to record metadata from approximately 1,400 cards. The students were provided with written procedures as well as a visual job aid to make the transfer of data from card to spreadsheet as clear as possible (see poster). Having the students enter codes directly from the cards without translating them with the key served to reduce the labor time per card and eliminate mistranslation errors. In addition, Excel functionality could be used to isolate invalid data in individual columns based on the valid value ranges for some columns.

The Metadata Librarian checked the students' output periodically throughout the project. Quality issues were minor and mostly typographical errors with number entry. The biggest hurdle was the handwritten text on the cards: in some cases handwriting was difficult to decipher, especially for proper names. Students were instructed to note entries that were difficult to decipher, so that the Metadata Librarian could examine the cards and do additional research as needed. A portion of the problem cards were completed by a paraprofessional from the Metadata Services Department after a student recorded the numeric coding.

Once the card metadata was captured, the Metadata Librarian split columns with multiple entries into individual columns. This resulted in multiple columns for several categories such as structures and technical notes. Once each column contained a single data point, another round of quality control was performed. The Metadata Librarian used conditional formatting to highlight invalid entries in each column. In some cases, a variety of invalid entries were searched for (e.g., letters and numbers outside of the valid range) and some spot checking was done against individual cards.

Following quality control, numeric codes were replaced by textual terms from the key column-by-column. Since each card might represent multiple slides, the Digitization Lab Manager de-duplicated entries on the spreadsheet by comparing it with the actual slides, indicating redundant slide numbers, or those for which we had no corresponding slide. The Metadata Librarian then further divided the document's rows into one per slide, removing entries for missing or redundant slides. The Metadata Librarian then crosswalked the spreadsheet data into the DC form and loaded it into the digital library software. The entire collection, including non-slide material, was processed and published in the CU Digital Library in time for the centenary of the Ludlow Massacre of 20 April, 1914, a watershed event in mining history and labor relations in the United States.

4. Conclusion

The Colorado Coal Project Collection, as it exists in the University of Colorado Boulder Archives, is a large, complex, and rich resource for researchers in mining and labor in the United States. Capturing and displaying the robust metadata that accompanied it proved an interesting and significant challenge, and served as a lesson in dealing with legacy metadata.

Acknowledgements

The author would like to acknowledge the hard work of everyone who contributed to this project, especially students Sarah Han, Rosalyn Wong, and Grace Zhong; library technicians Cynthia Hardey and Jane Zumwalt; Digitization Lab Manager Patrick Mulcrone; Digital Initiatives Librarian Holley Long; and original researcher and project sponsor Ron McMahan.

References

Dublin Core Metadata Initiative. (2012). *DCMI Metadata Terms.* Retrieved from http://dublincore.org/documents/dcmi-terms/

McCoy, Ralph E. (1965). *Computerized circulation work: a case study of the 357 Data Collection System.* Library Resources & Technical Services *9*(1), Winter 1965, 59-65.

Rabinow, Jacob. (1958). *Presently available tools for information retrieval.* Electrical Engineering *77*(6), June 1958, 494-498.

Smith, Donald A. & Peter L. Schnall. (1980). *Improved hypertension control using a surveillance system in a neighborhood health center.* Medical Care *18*(7), July 1980, 766-774.

Best Practice Demonstration: A Model and Roles of a Common Terminology to Improve Metadata Interoperability

(Boaz) Sunyoung Jin
University of Illinois at
Urbana-Champaign,
United States
sunjin@illinois.edu

Keywords: metadata; interoperability; common terminology; metadata model; MARC; (Q)DC

1. Introduction

Interoperability issues pose a barrier to sharing and exchanging information among digital libraries and repositories. This is due to the use of diverse metadata standards, and their different degrees of generality or specificity. This causes loss of information at all metadata model levels (e.g., schema, schema definition language, record, and repository) (Chan & Zeng, 2006) (Haslhofer & Klas, 2010, p. 19). As a possible solution for a long-term problem, historically argued standardization on a common communications format (Svenonius, 1983, p. 2), and a common command language or vocabulary (Lancaster & Smith, 1983, p. 21) are considered. A Common Terminology (CT), thus, is suggested as a bridge to various degrees' metadata standards to give uniformity for searching and to achieve metadata interoperability at multiple levels.

2. The Abstract Model and Roles of a Common Terminology (CT)

Based on DCMI abstract model (DCMI, 2013), an abstract model of CT is diagrammed in Figure 1. The definitions for terms in this extended abstract model are as follows:

- A Common Terminology is a set of Common Terms of element names in widely used metadata schemas such as MARC, MODS, DC and QDC.
- A Common Term is a property (element) or class.
- A property (sub-property) can be one kind of common element (field) or attribute (subfield) in two or more metadata schemas.

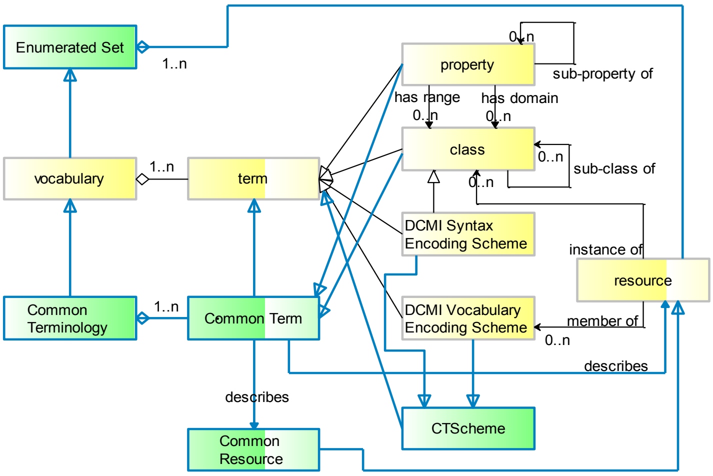


FIG. 1. The CT Abstract Model based on DCMI abstract model (DCMI, 2013)

The core role of CT is to encompass various metadata schemas allowing communities to use their own standards, while providing uniformity to searching. CT is a bridge of existing standards to maintain balance between different degrees of generality or specificity, minimizing loss of information at all metadata model levels. CT is to provide uniformity for search with CT union catalog and Linked Open Data connecting online accessible metadata records on the Web. CT, ultimately, is to provide a common standard way to achieve interoperability at multiple levels in order to share resources readily among many libraries, organizations, and governments.

3. The Developed CT to Improve Metadata Interoperability

Taking commonly used standards (MARC, MODS, DC, and QDC) as bases, CT has developed as a bridge across different generality and specificity levels. CT is selected to improve metadata interoperability at the schema, schema definition language, record, and repository model levels..

3.1. At the Schema Metadata Model Level

The developed CT (Jin, 2014) is a set of 12 Common Terms (properties), and 58 qualifiers (sub-properties) that specify and subdivide 12 properties in detail, with CTScheme. CTScheme is defined as a controlled set of values that are specific to CT. The development bases on crosswalks of Library of Congress (e.g., MARC from/to (Q)DC, etc.) (LC). The development is supported by usages of MARC tags and (Q)DC elements in 5 search interfaces and in actual metadata records of Harvard (MARC, 12 million records), UIUC (MARCXML, 10 million), and MIT (QDC, 20,000) through cooperation of three universities in the USA. The selected CT at the schema level is generalized common terms which maximize lexical and semantic interoperability, used over 50% usage in Harvard, WorldCat and UIUC metadata records; and used in all 5 search interfaces. 12 Common Terms are contributor, date, description, format, identifier, language, publisher, relation, rights, subject, title, and typeGenre. 58 qualifiers are on the project website.

3.2. At the Schema Language Definition Level

The generalized 12 Common Terms and 58 qualifiers are represented with XML schema (ct.xsd) and RDF schema (ct.rdf) with SKOS concepts (ctskos.rdf) to improve semantic interoperability.

3.3. At the Record Level

The performance of CT in achieving and improving metadata interoperability is presented through empirical evaluations with Harvard (MARC), MIT (QDC), and UIUC (MARCXML) records through cooperation of three universities. A conversion with Python language is designed to convert (Q)DC of MIT records to CT, and to measure transfer rate and lexical and semantic match rates. As a result of the conversion of mapping experiments, total transfer rate from (Q)DC of MIT to CT is 99.9%. Lexical and semantic match rates are 98.7% and 100%. Loss of information rate is extremely lower as 0.00463%. CT, thus, maximizes lexical and semantic interoperability reducing significantly the gaps of different degrees of generality or specificity. Finally, CT minimizes considerably loss of information at multiple levels.

3.4. At the Repository Level

As a next step, a prototype is planned to achieve and improve metadata interoperability at repository level. The prototype will build CT union catalog and Linked Open Data connecting 3 million online accessible records of Harvard (MARC), MIT (QDC) and UIUC (MARCXML) libraries providing a portal for them. The prototype will demonstrate a certain solution to build interoperability globally with CT among libraries or Well-Designed Digital Libraries all over the world that will consist of International Open Public Digital Library (Jin, 2014).

Conclusion

The Common Terminology (CT) has developed as a bridge across different generality and specificity levels such as MARC, MODS, DC, and QDC. CT minimizes considerably loss of information reducing the gaps among them. CT increases significantly accuracy in mappings showing high lexical and semantic match rates. The planned prototype will build CT union catalog and Linked Open Data connecting records of three universities on the Web, and provide a portal for Harvard, MIT and UIUC libraries. CT will give an assured solution to achieve and improve interoperability among university libraries and further among libraries and organizations to work together and share information reducing loss of information at multiple metadata levels.

Acknowledgements

This research project has been supported by the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign (UIUC). Thanks to Professors David Dubin and Linda Smith for their guidance on this project. We also appreciate the cooperation of Harvard, MIT, and UIUC in providing their metadata for the project.

References

Chan, Lois M., & Marcia L. Zeng. (2006, 06). Metadata Interoperability and Standardization – A Study of Methodology Part I. Achieving Interoperability at the Schema Level. D-Lib Magazine, Volume 12(Number 6).

DCMI. (2013). DCMI Abstract Model. Retrieved from Dublin Core Metadata Initiative: http://dublincore.org/documents/abstract-model/

Haslhofer, Bernhard, & Wolfgang Klas. (2010). A Survey of Techniques for Achieving Metadata Interoperability. ACM Comput. Surv., 42(2).

Jin, (Boaz) Sunyoung. (2014). A Model and Roles of a Common Terminology to Improve Metadata Interoperability. Illinois Digital Environment for Access to Learning and Scholarship (IDEALS). Retrieved from http://hdl.handle.net/2142/50100

Jin, (Boaz) Sunyoung. (2014). International Open Public Digital Library (IOPDL): A Proposal for the Future. Illinois Digital Environment for Access to Learning and Scholarship (IDEALS). Retrieved from http://hdl.handle.net/2142/50101

Lancaster, F. Wilfrid, & Linda Smith. (1983). Compatibility Issues Affecting Information Systems and Services. General Information Programme and UNISIST.

LC. (n.d.). Conversions. Retrieved from Metadata Object Description Schema (MODS): http://www.loc.gov/standards/mods/mods-conversions.html

Svenonius, Elaine. (1983). Compatibility of Retrieval Languages: Introduction to a Forum. Int. Classif, 10(No.1), 2-4.

Best Practice Poster: Converting Personal Comic Book Collection Records to Linked Data

Sean Petiya
Kent State University, USA
spetiya1@kent.edu

Keywords: Linked Data; comic books; graphic novels; ontologies; metadata; usability.

1. Introduction

The Comic Book Ontology (CBO) is a metadata vocabulary currently in development for the description of comic books and comic book collections. The vocabulary is part of a larger, ongoing research project exploring the design and exchange of data about comic books and graphic novels. The goal of the project is to produce a series of usable schemata and tools for the many participants in the often complex universe of comic books, which includes publishers, collectors, and libraries, among many others. The long-term objectives of the project include addressing the needs and overlapping roles of each user group through designated application profiles. Recognizing that all groups involved will have different needs, goals, and concerns, the base for each of these user application profiles is a much simpler set of elements required to first uniquely identify a resource. The intention of this core set of elements is to lower the difficulty in implementing the vocabulary and enhance the overall understandability of the ontology. The core application profile has been modeled from common elements found in the data of comic book collectors, a community of users largely responsible for the preservation of the medium, which has historically been underrepresented in knowledge institutions.

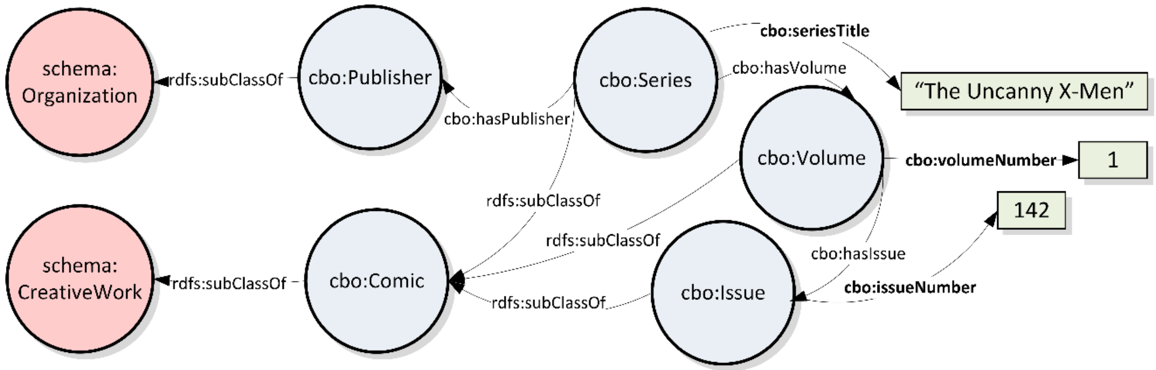


FIG 1. Core concepts in the Comic Book Ontology (CBO).

This poster describes progress on the Comic Book Ontology (CBO) by presenting a diagram illustrating current components of the model (FIG. 1), and outlines the methodology and rationale for producing a core application profile. Additionally, it presents a workflow illustrating how the core set of elements is used to map user data to the vocabulary and generate RDF/XML records through an automated process. Community data is commonly contained in spreadsheets, or made available as CSV, and a workflow is described for both the preparation and conversion of that data, as well as its connection to existing Linked Open Data (LOD) resources.

2. Background

The recent success of Marvel's *Guardians of the Galaxy* at the box-office highlights the dominance of the superhero movie in popular culture, and interest in the genre is only likely to continue with future films planned featuring familiar icons like Batman, Superman, and Spider-Man. However, before these characters and stories made it to movie screens, they first appeared in periodical comic books on newsstands, where they then made it into the homes and collections of many generations of readers around the world. In addition to appearing in library special collections and archives, like the Comic Art Collection of the Michigan State University Library composed of over 200,000 items (comics.lib.msu.edu), the comic book is also collected by the Library of Congress (LOC) and the institution's Comic Book Collection contains over 120,000 comic issues (LOC, 2013). While the efforts of these institutions are significant, parallel activities occur daily in the homes of many comic book collectors (Serchay, 1998). Passion and dedication to the hobby on the part of both collectors and professionals has produced numerous research projects and efforts dedicated to the comic book. Notable projects in this area include the *Grand Comics Database* (GCD), an international effort to index all comic books published worldwide (gcd.org), and *Comichron: The Comic Book Chronicles*, a research project collecting comic book sales and circulation data (comichron.com), among many other related endeavors that can be found in the *Comics Research Bibliography* (Rhode & Bullough, 2009). The Comic Book Ontology (CBO) represents an effort to bring greater bibliographic control, representation, and visibility to the endeavors of many writers, artists, researchers, and collectors who have contributed to the preservation and proliferation of the medium.

3. Application Profile and Workflow

The comic book is a complex object that can be viewed as a bibliographic resource, collection item, and art object, with its contents telling part of the story in an ongoing narrative that can span multiple issues, volumes, and series titles, all of which compose a detailed, fictional universe. In addition to the complexities of the objects themselves, the domain's many participants, including libraries and archives, each produce data of various degrees of quality and control, while following different standards and practices. However, shared entities and elements found in the data formulate a core model that can represent a simplified view of this complex world.

The methodology for producing the core application profile involved aligning components of the Comic Book Ontology (CBO) to a WEMI model. The WEMI model produces a view of the core elements at various levels of description, up to a specific, physical copy in a comic book collection. Extending the exchange of knowledge to the collector using Linked Data enables a passionate and dedicated segment of the user population to participate in the ecosystem, not just at the item-level, but at all levels of resource description potentially expanding the "global graph" of RDF statements describing comic works, creators, and collections. However, in order to participate successfully, users require a simple, clear process for the preparation and conversion of their data. This workflow involves: (1) mapping existing data to CBO terms, (2) converting data to qualified RDF/XML, and (3) automatically replacing values with LOD URIs. The automated conversion process is achieved through an online tool, or a script that can be run locally. Experienced users can modify, rewrite, or create their own script, and expand on the selection of LOD resources linked in the resulting dataset.

4. Summary

The Comic Book Ontology (CBO) seeks to provide the tools through which collectors, researchers, and libraries can share information about their individual collections and better combine and exchange knowledge in a Linked Data environment. In order to improve the usability of the ontology, a core application profile has been developed. A basic workflow describes using this profile to guide the preparation and mapping of existing data to CBO elements, and the automated conversion of that data to qualified RDF/XML containing Linked

Data URIs for common values. The core application profile will form the base of additional profiles that will address the needs of other user groups as the ontology expands. The vocabulary is made available at comicmeta.org, which functions as a repository for the ontology as well as all related schemata, tools, and utilities.

References

Library of Congress. (2013, April). Comic book collection. Retrieved from http://www.loc.gov/rr/news/coll/049.html

Rhode, Michael, and John Bullough. (2009). Comics research bibliography. Retrieved from http://homepages.rpi.edu/~bulloj/comxbib.html

Serchay, David. S. (1998). Comic book collectors: The serials librarians of the home. Serials Review, 24(1), 57-70. doi:10.1016/S0098-7913(99)80103-8

Best Practice Poster: Making Vendor-Generated Metadata Work for Archival Collections Using VRA and Python

Carolyn Hansen
University of Cincinnati
United States
carolyn.hansen@uc.edu

Sean Crowe
University of Cincinnati
United States
sean.crowe@uc.edu

Keywords: vendor-generated metadata; metadata mapping; archival description; Dublin Core; VRA; Python

1. Introduction

Although cataloging cultural resources requires a greater level of descriptive granularity than standard library materials, metadata for digital collections is often generated by non-specialists. This can lead to significant problems with metadata accuracy and consistency, causing breakdown of authority control, high incidence of false positives in searching, and impeded access to materials. The purpose of this poster is to illustrate a successful workflow for improving vendor-generated metadata for a large digital collection of archival materials by converting the metadata from the Dublin Core standard to the VRA standard using the scripting language Python.

2. Background

The University of Cincinnati Libraries (UCL) contracted with a vendor to scan and generate metadata for the Cincinnati Subway and Street Improvements Collection. Consisting of photographs and documents related to the construction of the unfinished Cincinnati Subway system and street improvements throughout the city, the collection is a unique resource documenting early 20th century transportation, urban planning, and social history. Following the initial load of approximately 9,000 scanned images and associated Dublin Core metadata records into the shared OhioLINK Digital Repository Center, librarians Sean Crowe and Carolyn Hansen were charged with converting the metadata to the VRA standard, improving metadata quality, and loading the collection into the University's Luna image repository. Carolyn Hansen brought metadata standard expertise and Sean Crowe provided technical and scripting skills to the project.

3. Implementation

The planning and specifications for the contract scanning project were conducted by UCL's Digital Projects Repositories Department, and did not include input from UCL's Content Services Division, in which the authors work. As a result, the project workflow began with an assessment phase, which involved researching the initial scanning project, assessing the vendor-generated metadata, and gathering domain-specific information about the original physical format of the materials. A metadata map was created to record decisions about field equivalents between Dublin Core and VRA, controlled vocabulary usage, improvement of vendor-generated metadata, and addition of VRA-specific fields to describe original materials and digital surrogates.

These decisions were then encoded into a Python script. The Python script incorporated a custom class to parse and process the metadata in CSV format. In addition to coding the field conversions and formatting field contents based on the metadata map, the script ran several validation processes on the input and output metadata files. Finally, a function was added to the script to link records to image files by unique identifier. Coding the script comprised a considerable portion of the project timeline though the script run-time was negligible.

4. Challenges

Project implementation involved a number of challenges. In terms of metadata mapping, moving from a less robust standard like Dublin Core to a very robust standard like VRA required strategic decisions. Since VRA provides the opportunity for highly-detailed descriptive metadata, it is necessary to look at the metadata with a strong editorial eye in order to balance detailed description with project time constraints and vendor-created metadata of varying quality. In order to accomplish this, a baseline for acceptable metadata was created, detailing changes to vendor-created metadata as well as who would be responsible for metadata enrichment. For example, errors in access points from controlled vocabularies such as LCNAF or LCSH headings would be corrected by Content Services faculty, but additional subject analysis would be provided by curators at a later stage in the project. The metadata quality baseline was also applied to controlled vocabulary usage. For example, when working with detailed vocabularies like the Getty Research Institutes' Art & Architecture Thesaurus, it was important to balance the level of descriptive granularity with vocabulary that was understandable to users and applicable to a wide range of materials.

Additionally, local practices regarding archival materials presented unique challenges to the project. Specifically, university archivists at UCL preferred that the structure of the digital collection should replicate the physical archive, including record order and collection level titles for item records. As a result, titles without description of the image content such as "Rapid Transit Photographs -- Box 17, Folder 22 (September 21, 1922 - October 24, 1922) -- negative, 1922-09-28, 9:42 A.M." were used. These titles offer little descriptive content and create greater reliance on subject searching. Further work needs to be done to make the collection searchable based on the content of the image. Lastly, geographic coordinates, included in some of the records, enrich the collection and should be added where possible.

5. Conclusions/Results

Since the collection was posted in Fall 2013, it has received over 17,000 unique page-views in the Luna Repository. This project serves as a template for future shared, interdepartmental projects. Further collaboration is certain as traditional Library Technical Services operations evolve to support local and unique digital content, including research data, archival material, and beyond.

Acknowledgements

Linda Newman, Head, University of Cincinnati Libraries Digital Content & Repositories Dept. and Elna Saxton, Head, University of Cincinnati Libraries Content Services Division.

References

Crowe, Sean and Carolyn Hansen (2014). DC_to_VRA. In GitHub. Retrieved from https://github.com/crowesn/DC_to_VRA.

University of Cincinnati Libraries (2014). Cincinnati Subway and Street Improvements, 1916-1955. Retrieved from http://digital.libraries.uc.edu/subway/.

University of Cincinnati Libraries (n.d.). LUNA Digital Repository. Retrieved from http://digproj.libraries.uc.edu:8180/luna/servlet/univcincin~42~42>.

Best Practice Poster: A Library Catalog REST API Framework

Jason Thomale
University of North Texas, United States
jason.thomale@unt.edu

William Hicks
University of North Texas, United States
william.hicks@unt.edu

Keywords: library catalog metadata; MARC; Machine Readable Cataloging; REST; Representational State Transfer; integrated library systems; API; application programming interface;

Abstract

Within the archipelago of cultural memory data, library catalogs and systems still comprise some of the most isolated and least penetrable desert islands. Although the library world has made significant strides over the past decade to open its metadata, many individual libraries remain at the mercy of their ILS vendors to implement open protocols, standards, and APIs. At the University of North Texas Libraries, we have been developing a REST API framework for exposing our catalog and ILS metadata, taking our first steps toward breaking our data off this particular island.

Catalog resources that we've modeled so far include bibliographic records (modified from MARC), item-level records, branch location records, item type records, and item status records. We are also working on resources that support a shelf-list browser application, which mix user-supplied data with item and bibliographic metadata and demonstrate a real-world use for the API.

Our framework is not merely an API for our particular ILS. Rather, we are developing a toolset to allow us to extract and re-model our ILS data—to use data derived from our ILS but not necessarily to adhere to ILS data models—and expose the data as RESTful, linked resources. Although our initial efforts have focused on modeling resources that do closely align with ILS entities, future development will include extended models for work- and identity-related resources and possibly extending our APIs to expose linked data (using, e.g., JSON-LD).

Best practices in this area, exposing ILS metadata as RESTful resources, are hard to come by. Given the mixture of metadata practitioners, systems-oriented individuals, and web-oriented individuals that the Dublin Core Metadata Initiative (DCMI) conferences tend to attract, we hope that presenting a poster about the project in the Best Practices track might allow us to connect with others with whom we might dialog. Ultimately, we believe an exchange of information about our project so far—our approach and practices—would be valuable to us and to others in the DCMI community.

Best Practice Poster:
Building the Bridge: Collaboration between Technical Services and Special Collections

Susan Matveyeva
Wichita State University Libraries
Susan.Matveyeva@wichita.edu

Lizzy Walker
Wichita State University Libraries
Lizzy.Walker@wichita.edu

Keywords: departmental collaboration; standards; best practices; CONTENTdm; metadata

1. Introduction

At Wichita State University Ablah Library, members of Technical Services and Special Collections began collaborating on a mass digitization project to increase visibility and accessibility of Special Collections holdings, and to digitally preserve brittle rare materials. Both departments scan collections, create metadata, and upload materials into CONTENTdm. The departments overcame challenges regarding the project, such as limited collaboration between the departments, poor communication, minimal metadata, and differences in quality control expectations.

2. Challenges

Differences in philosophy between Technical Services and Special Collections presented the first challenge. Special Collections was concerned with securing their collections and felt librarians did not share this concern. They also emphasized the collections over users' needs. Since their practice was boutique style treatment for items, less attention was paid to productivity, and keeping current with changing cataloging standards. Special Collections were concerned Technical Services were unfamiliar with archival practices. Technical Services goal was to operate with the end user in mind. To that end, the adoption of RDA as well as OCLC's *Best Practices for CONTENTdm and other OAI-PMH Compliant Repositories* (2013) in metadata creation reflected this focus. They also had a provider/client relationship at first. Technical Services also felt Special Collections were not familiar with the standards and practices the cataloguers used, nor with Technical Services' production environment.

Poor communication presented another challenge. When Technical Services completed a collection, it was returned to Special Collections with expectation of rapid feedback. With no information forthcoming, Technical Service operated as though there were no problems. As a result, they completed six collections by the time Special Collections sent feedback. Some corrections came from incorrect information from old finding aids. Additionally, Technical Services and Special Collections each had internal control processes, but no shared criteria existed for gauging quality.

Common metadata standards did not exist between the departments. Technical Services operated in a production environment based on collaboration and cooperation. Special Collections operated in an isolated environment with emphasis on unique description, locally created metadata, and traditional archival standards. They also did not have a dedicated cataloger on staff. The uniqueness of uncontrolled vocabulary metadata versus controlled vocabulary for interoperability allowed for much constructive debate. The goal in regard to metadata was to get Special Collections and Technical Services using the same standards.

The departments had different approaches to metadata. Special Collections focused mainly on the descriptive metadata in a human-readable format. Technical Services was interested in the addition of administrative and technical metadata, and kept in mind the current Web environment and machine-readable representation of information.

3. Method and Results

Building trust between the departments involved many facets. Staff from both departments created a metadata group responsible for creating metadata templates for manuscripts and printed materials. Investigation of standards and best practices, creation of data dictionaries, and mapping templates were only a few of the topics focused on by this subcommittee. The group developed minimal and core level metadata templates for published and unpublished materials based on common standards for rare books and manuscripts using OCLC's *Best Practices*. The templates focused on access to collections, future migration, and preservation. Technical Services accommodated unique needs of Special Collections while working on the creation of shared workflows and metadata templates. Special Collections responded positively to the processes and also recommended changes based on their needs. Multiple revisions to the templates were required to accommodate both departments.

Quality control quickly became a priority. With administration support, the departments implemented pre-planning meetings where the departments discuss specific collections. This includes the level of metadata Special Collections and Technical Services selects for a collection. Levels of quality control are also present throughout the process. Multiple people handle the scans and metadata in terms of viewing and uploading, as well as the final review. There are also pre-planning meeting forms, scan inventory worksheets, a metadata cheat sheet for the catalogers, and workflow checklists.

Technical Services introduced DC mapping in CONTENTdm, as well as an enhanced production environment to Special Collections that previously performed boutique treatment of materials. Likewise, Special Collections communicated their specific needs so that we gained an understanding of expectations from Special Collections, which the metadata group kept in mind when creating the templates. Special Collections' willingness to work with the OCLC Best Practices, as well as RDA was a real leap for the department in terms of opening up their collections to a worldwide audience.

3. Next Steps

Next steps include appointing a project manager who will lead each project from beginning to end. A metadata checklist is being created to aid the catalogers in reviewing their peers' work. Creation of local controlled vocabularies will also be a future project.

4. Conclusion

This has been a positive collaborative experience for both departments. Bringing expertise of catalogers and uniqueness of Special Collections together has helped them to be less isolated. The implementation of metadata and cataloging standards creates a layer of interoperability, and increases the potential of users finding unique materials. Additionally, the departments have a new working relationship that will hopefully continue in the future.

Acknowledgements

The authors would like to thank Wichita State University Libraries Special Collections and University Archives for their collaboration and partnership in creating this project.

References

OCLC. (2013). Best practices for CONTENTdm and other OAI-PMH compliant repositories: Creating sharable metadata. Retrieved June 01, 2014, from http://www.oclc.org/content/dam/support/wcdigitalcollectiongateway/MetadataBestPractices.pdf

Best Practice Poster: Best Practices for Complex Diacritics Handling in CONTENTdm

Jason W. Dean
University of Arkansas
Libraries, USA
jwdean@uark.edu

Deborah E. Kulczak
University of Arkansas
Libraries, USA
dkulczak@uark.edu

Keywords: CONTENTdm, diacritical marks, indexing, UTF, encoding

In order to ensure the best possible access for materials held by libraries and archives, these institutions must employ special accent and punctuation marks when transcribing or transliterating languages other than English. These marks are called diacritical marks by the library community. Their use in MARC cataloging is widespread, as is their use in library catalogs. However, users of digital content management systems (CMS), such as CONTENTdm encounter difficulty in ensuring appropriate diacritical marks are read by the CMS when metadata is imported or migrated into such a system. These problems further compound searching issues for the user, as noted by Bar-Ilan and Gutman (2005) in the *Journal of Information Science*. Little literature and few instructions exist to assist users in working with these diacritical marks. However, some pertinent literature exists on the subject.

In Hongyan Jing's essay for an IEEE symposium on speech synthesis (2002), the author in discussing Italian highlights the ubiquity of all types of diacritical marks. His work states that of 445,626 entries in a dictionary, 4.9% of these entries include a diacritical mark. Though the number is not high, for libraries and archives this number represents a barrier to access and description that must be overcome. Tull and Straley's article in *Library Hi Tech* (2003) covers the issues presented in sorting and searching in relation to diacritical marks. Most literature discusses formatting text in UTF-8 or similar UTF standards however some literature discusses the use of ASCII. This poster focuses on the use of UTF-8, which is required by CONTENTdm to ingest diacritical marks correctly.

The research and work behind this poster came largely from a recently completed project at the University of Arkansas Libraries that dealt with metadata and items in a plethora of languages, from English and French to Quapaw, many of which required the use of unusual diacritical marks. The authors were responsible for the ingestion of metadata into CONTENTdm and encountered several issues with complex diacritical marks presented by the disparate languages in this project. What follows is the procedure arrived at and now codified in a metadata "cookbook."

The handling of these diacritical marks was primarily in three areas: controlled vocabularies in CONTENTdm, transcripts, and loading metadata spreadsheets.

Creating and importing a controlled vocabulary list is most easily done in Notepad++ and encoded as "UTF-8 without BOM." Using these settings, diacritical marks ingested into CONTENTdm will be maintained using this encoding setting and following the CONTENTdm instructions for loading a controlled vocabulary.

Transcripts are best handled using a similar procedure. Transcripts are created in Notepad++, and saved as "UTF-8 without BOM" as the encoding setting. However, some transcripts might be loaded at the same time as metadata in a spreadsheet. In this case, if the spreadsheet is created in Excel, the user must use the "Arial Unicode MS" font for data entry. When data entry is complete, use the Save As command to save the spreadsheet as a tab-delimited text file. In the Save As dialog box, select "Unicode Text" from the "Save as type" menu. After selecting "Unicode Text", select the "Tools" box to the left of the "Save" button. Select the "Encoding" tab in the "Web Options" dialog box. In the "Save this document as" box, select "Unicode (UTF-8)" from the drop-down menu. Select "OK" then "Save" in the Save As menu.

Metadata spreadsheets and tab-delimited files present a similar set of challenges for diacritical marks loading into CONTENTdm. In this case, if the spreadsheet is created in Excel, the user must use the "Arial Unicode MS" font for data entry. When data entry is complete, use the Save As command to save the spreadsheet as a tab-delimited text file. In the Save As dialog box, select "Unicode Text" from the "Save as type" menu. After selecting "Unicode Text", select the "Tools" box to the left of the "Save" button. Select the "Encoding" tab in the "Web Options" dialog box. In the "Save this document as" box, select "Unicode (UTF-8)" from the drop-down menu. Select "OK" then "Save" in the Save As menu.

References

Bar-Ilan, Judit, and Tatyana Gutman. (2005). How do search engines respond to some non-English queries? Journal of Information Science. 31(1), 2005, 13-28.

Jing, Hongyan. (2002). Identifying accents in Italian text: a preprocessing step in TTS. Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002, 151-154.

Tull, Laura, and Dona Straley. (2003). Unicode: Support for multiple languages at the Ohio State University Libraries. Library Hi Tech. 21(4), 2003, 440-450.

Best Practice Demonstration: Ecco!: A Linked Open Data Service for Collaborative Named Entity Resolution

Matthew Miller
New York Public Library,
NYPL Labs,
United States
matthewmiller@nypl.org

M. Cristina Pattuelli
Pratt Institute, School of
Information and Library
Science,
United States
mpattuel@pratt.edu

Keywords: named entity resolution; Linked Open Data (LOD); Ecco!;

Abstract

This demo proposal presents Ecco!, a Linked Open Data (LOD) application for entity resolution. Specifically, Ecco! is designed to disambiguate and reconcile named entities with URIs from authoritative sources. Technically, Ecco! creates a wrapper around LOD APIs of suitable datasets such as VIAF and Freebase to retrieve data useful for supporting entity matching. The system automatically ranks and groups the results into different clusters according to various confidence levels – from exact matches to one to many or no matches. The quality of the data output can be further refined through human disambiguation consisting of validating a match or identifying the correct URI when multiple matches are possible.

Ecco! is designed to enable users to quickly and easily contribute to this curation process. The system provides an intuitive user interface that supports a collaborative workflow where a community can work together in a distributed and incremental way. The combination of automated matching plus human curation has the potential to produce a superior quality of data, not currently achievable through traditional methods.

This application works alongside existing legacy systems and data sources through an import and export workflow. Extracts generated from a legacy system or data source are enriched through Ecco! and then looped back to update the originating source. Ecco! intends to address the well-known "bucket names" problem that occurs when legacy data has accumulated and contains a mix of heterogeneous names derived from different authorities (e.g., LC/NAF, ULAN, etc.) as well as locally defined terms.

Ecco! is a node.js application that anyone can download and run on their local system. There is no need for a server installation, but it could be installed on a server to allow for the collaboration of an unlimited number of participants. Ecco! has the capacity to work with LOD APIs in a modular way. While the demo version will specifically leverage VIAF and Freebase, any API plugin could be virtually written for it. Also, while in the current release the application will be centered on persons and organizations, other types of entities including geographic locations, events, topics, etc. could be also handled by the system.

Even though Ecco! was developed as part of the Linked Jazz project,[1] it is domain-agnostic and thus not tied to any specific context of use. The demonstration includes different scenarios showing a series of use cases. Results from a first round of testing will also be shared.

Data quality poses a daunting challenge in Linked Open Data development and requires the creation and adoption of new methods and tools to promote accuracy and consistency of data. Ecco! includes a series of innovative features that make it uniquely flexibility and easy to use.

[1] http://linkedjazz.org

Most notably, this system lowers the barrier for non-programmers who want to actively contribute to the production of high quality linked data through a user-friendly and collaborative platform.

Best Practice Poster: Wikipedia-based Extraction of Lightweight Ontologies for Concept Level Annotation

Elshaimaa Ali
University of Louisiana at Lafayette, USA
eea7236@Louisiana.edu

Michael Lauruhn
Elsevier Labs, USA
m.lauruhn@elsevier.com

Keywords: Wikipedia; text mining; annotation; semantic annotation; lightweight ontologies

Abstract

This poster describes a project under development. We propose a framework for automating the construction of lightweight ontologies for semantic annotations. Lightweight ontology is defined as the ontology that does not have to include all the components expressed with formal languages such as concept taxonomies, formal axioms, disjoint and exhaustive decomposition of concepts. (Giunchiglia and Zaihrayeu 2009). However, manual enhancement of the ontology through the addition of axioms, rules, disjoint sets, etc., is possible for future reasoning purposes. The purpose behind this research is to evaluate possible means for efficiently annotating domain-specific content using open ontology sources.

When considering building ontologies for annotations in any domain, we follow the process of ontology learning in (Stelios 2006) which are: acquisition of the relevant terminology, identification of synonym terms / linguistic variants, formation of concepts, hierarchical organization of the concepts (concept hierarchy), learning of relations, properties or attributes, together with the appropriate domain and range, hierarchical organization of the relations (relation hierarchy), instantiation of axiom schemata, definition of arbitrary axioms, and ontology evaluation. Since we are looking for a lightweight ontology, we only consider a subset of these tasks, which are the acquisition of domain terminologies, generating concept hierarchies, learning relations and properties, and ontology evaluation.

When developing the framework modules we base most of our knowledge base on the structure of the Wikipedia, which represents the hierarchical links between categories and links between pages, in addition to specific sections of the content. To ensure machine readability and interoperability, ontologies have to be explicit to make an annotation publicly accessible, formal to make an annotation publicly agreeable, and unambiguous to make an annotation publicly identifiable (Ding 2006). An important aspect in order to achieve explicitly, formality and unambiguity of the developed ontology, is to define an annotation schema that allows the ontologies to be reused and be part of linked data.[1] We designed our schema based on annotation elements already defined in the Dublin core standards[2] and we used the DBpedia[3] annotation elements for defining named entities. We are also introducing new elements for annotating concepts and defining the context (domain knowledge) in which the concept exists.

The main tasks for this framework are: extracting domain concepts and terms, measuring relatedness between domain terms, defining boundaries of subdomains using concept clustering and extracting relations, and defining named entities within each subdomain.

The following figure is an abstract explanation to the modules of the proposed framework.

[1] http://linkeddata.org/

[2] http://dublincore.org/documents/usageguide/

[3] http://mappings.dbpedia.org/server/ontology/classes/

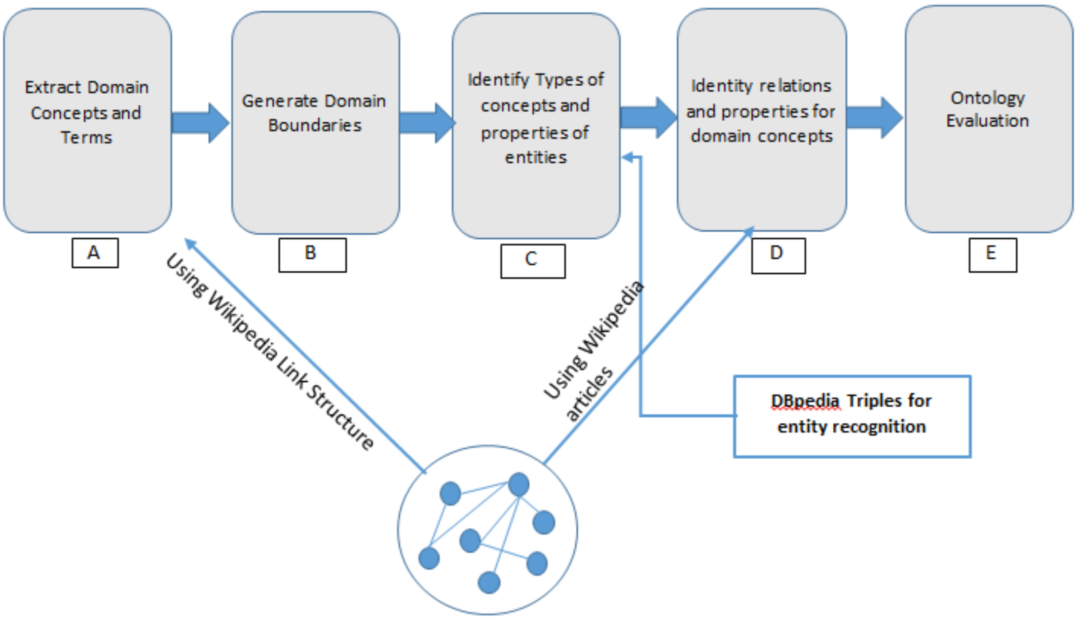

Extract Domain Concepts and Terms
A
Generate Domain Boundaries
B
Identify Types of concepts and properties of entities
C
Identity relations and properties for domain concepts
D
Ontology Evaluation
E
Using Wikipedia Link Structure
Using Wikipedia articles
DBpedia Triples for entity recognition

FIG. 1. Framework modules.

We start by defining a pool of domain terms and concepts that needs to be modeled for the domain ontology, but if this is not the case then we build Module A, where we start with relevant domain concepts and consider them as seed concepts for the ontology. Then we expand our concepts space using the Wikipedia link structure.

In Module B, we generate the domain and subdomain boundaries by computing relatedness between domain terms extracted in the first phase. Then we build a similarity matrix that models the relatedness between the extracted concepts. We developed a relatedness measure that relies on the degree of connectivity between two concepts in the Wikipedia graph. We then use hierarchical clustering (Diday and Simon 1980) to create subdomain boundaries. In Module C, we classify the generated named entities and concepts into wiki concepts and named wiki entities according to the description of the annotation schema. We will use the DBpedia triples for named entity recognition. In Module D we extract concept hierarchies and concept – concept relations by analyzing sections of Wikipedia articles. We will use openNLP to parse and extract relations defined in sections like the introductory sections in the Wikipedia page that defines the concept, in addition to exploring the category graph for the Wikipedia. OpenNLP has been successfully used for extracting relations for ontology enrichment in (Barkschat 2014). For Module E, we will evaluate the extracted ontology by comparing it to some of the mature existing ground truth like predefined domain ontologies or even topic maps that is created by a domain expert, and will use manual and expert evaluations.

Acknowledgements

Research and development of the project described in this poster and abstract is part of the LinkedUp Project, a European Commission FP7 Support Action.

References

Barkschat, K. (2014). Semantic Information Extraction on Domain Specific Data Sheets. The Semantic Web: Trends and Challenges, Springer: 864-873.

Diday, E. and J. Simon (1980). Clustering analysis. Digital Pattern Recognition, Springer: 47-94.

Ding, Y., Embley, D W . (2006). Using Data-Extraction Ontologies to Foster Automating Semantic Annotation. ICDE.

Giunchiglia, F. and I. Zaihrayeu (2009). Lightweight ontologies. Encyclopedia of Database Systems, Springer: 1613-1619.

Stelios, K., Dimitris, A. (2006). "Consensus Building in Collaborative Ontology Engineering Processes." Journal of Universal Knowledge Management vol. 1 (no. 3): 199-216.

Best Practice Poster: How to Build a Local Thesaurus

Robert H. Estep
Fondren Library, Rice University
United States
estep@rice.edu

Keywords: LCNAF; local headings; thesaurus construction.

1. First Steps

One of the first steps taken in preparing for participation in the Rice Historical Images Project was the assembling of a skeletal structure of terms specific to both Rice University and its earlier incarnation as the William M. Rice Institute.

The primary research tools used were archival maps and blueprints, newspaper accounts contemporaneous with the University's building schedule, campus telephone directories, and online entries (including Wikipedia). Alphabetical lists of building names, and names of University departments and schools, were cross-referenced for name changes effective during the University's history, and checked against the corporate LCNAF. The more complex internal inconsistencies were noted for future fiddling and/or resolution.

2. People, Places, and Things

It was clear from the start that an unusually large number of LOCAL headings would need to be constructed, the bulk of these in the form of corporate headings for departments and organizations, such as the Rice MOB (or Marching Owl Band), university buildings and structures, as well as headings related to the city of Houston. In some cases additional research was required, one example being that of the historic Rice Hotel in downtown Houston, for which construction and demolition dates were included in the heading.

Once the project commenced a large number of personal name headings were also required, for faculty, staff, students, members of the Houston business community, city and state politicians, and others. Additionally, in the case of University faculty with existing LCNAF entries, a second LOCAL heading was included with the parenthetical "(Faculty)" made explicit. Regional resources, often in the form of obituaries or published tributes and Festschrifts, were scoured for relevant dates, middle initials and names, and other information.

The smallest number of LOCAL headings were reserved for 'things': in other words, for events or activities which were unique to the history of Rice as an institution. The most prominent of these were the "May Fete" and the "Spring Rondolet" (both dance events), as well as the thematically adventurous "Archi-Arts Ball", not to mention the yearly "Beer Bike cycling event". In addition to events such as these, certain architectural features were given headings because of the frequency and prominence of their appearances in the images, for instance, the large central "Sallyport" which can be seen from Main Street and which ushers visitors and members of the Rice community alike into the large central "Quad", bounded on one end by Fondren Library.

3. The Desire for Consistency

The use of LOCAL headings gave us the latitude to delve deeply into the metadata description of each image, but every attempt was made to model these headings upon valid forms already in LC, whether the LOCAL heading was for a building name, a student organization, or a member of the faculty.

We quickly learned that bannering consistency, as one of the most important qualities of the thesaurus would demand patience and flexibility, as new image-types or additional archival information made our earlier entries obsolete or overly unique.

4. Looking Forward, Looking Back

Building a progressive thesaurus which is both a melding of valid LC headings and LOCAL headings requires the flexibility of being able to return from time to time, sometimes to tweak, other times to undo earlier work and start from scratch. But the pattern we have found is that each return is both easier and shorter, as we digest the lessons of embarking on a project which extends both into the past via the images themselves, and into the future as the life of the University, its teachers and its students, continues to be documented.

5. Illustrative Aspects

The poster will feature graphics in the form of sample images from the project collection itself, as well as screen shots of the existing Thesaurus.

Best Practice Poster: Designing an Archaeology Database: Mapping Field Notes to Archival Metadata

Ann Ellis
Austin State University Library
Stephen F. Austin State University
United States
aellis@sfasu.edu

Keywords: archaeology; field notes; archaeological artifacts

Abstract

The Stephen F. Austin State University Center for Digital Scholarship and Center for Regional Heritage Research engaged in a collaborative project to design and implement a database collection in a digital archive that would accommodate images, data and text related to archaeological artifacts located in East Texas. There were challenges in creating metadata profiles that could effectively manage, retrieve and display the disparate data in multiple discovery platforms.

The poster illustrates the steps that were taken to map field notes into useful archival metadata. Using original notes and field record information a preliminary data dictionary was created. After collaborative edits and revisions were made, a comprehensive data dictionary was designed to represent the materials in the collection. From this, a profile was configured in the digital archive platform to allow for upload of the metadata and images, and for discovery and display of the archaeological artifacts and related works.

Best Practice Poster: Utilizing Drupal for the Implementation of a Dublin Core-Based Data Catalog

Lisa Federer
National Institutes of Health Library,
United States
lisa.federer@nih.gov

Keywords: NIH; National Institutes of Health; data catalog; Drupal; content management system.

1. Objective

To create a data catalog suitable for use within the context of biomedical and health sciences research. The ideal catalog would allow researchers to easily describe their data using Dublin Core Metadata Terms and subject-appropriate controlled vocabularies, as well as provide search and browse capabilities for end users to enable data discovery and facilitate re-use.

2. Setting

The National Institutes of Health (NIH) Library serves the community of NIH intramural researchers, which includes over 1,200 principal investigators and 4,000 postdoctoral fellows conducting basic, translational, and clinical research on its primary campus in Bethesda, MD, and several satellite campuses.

3. Methods

Drupal, a free and open-source content management system, was utilized as a framework for a data catalog using the Dublin Core Metadata Terms. Using the Structure function within Drupal, the research data informationist at the NIH Library constructed a pilot system that utilized Dublin Core Metadata schema and relevant biomedical taxonomies. This pilot system can be adapted to the needs of a variety of basic, translational, and clinical research applications.

4. Results

The pilot system is currently undergoing testing with researchers within the NIH intramural community. Results will be available by the time of the DCMI 2014 conference.

5. Conclusions

A data catalog that utilizes an extensible metadata schema like Dublin Core and an open-source framework like Drupal provides users a powerful yet uncomplicated method for describing their data.

5. Implications

As funders and publishers increasingly require data sharing, researchers will need simple, intuitive methods for describing their data. Open-source systems like Drupal and extensible metadata schema like Dublin Core will likely play a large role in data description, thus making data more discoverable and facilitating data re-use.

Best Practice Poster: PunkCore: Developing an Application Profile for the Culture of Punk

Joelen Pastva
University of Illinois at
Chicago, United States
jpastva@uic.edu

Valerie Harris
University of Illinois at
Chicago, United States
val66@uic.edu

Keywords: DCAP; Punk culture; Punk music; domain model; functional requirements; genre vocabulary;

Abstract

PunkCore is a Dublin Core Application Profile (DCAP) for the description of the culture of Punk, including its music, its places, its fashions, its artistic expression through film and art, and its artifacts such as fliers, patches, buttons, and other ephemera. The structure of PunkCore is designed to be simple enough for non-experts yet specific enough to meet the needs of information professionals and to capture the unique qualities of materials classified as Punk. In the interest of interoperability and adoptability, PunkCore is drawn from existing metadata schema, and the development of PunkCore is intended to be open and collaborative to appeal to the entire Punk community. Our poster illustrates the initial development of the PunkCore standard and outlines future plans to bring PunkCore to the community.

The PunkCore DCAP is in its first phase of development, which follows Singapore Framework stages 1 and 2, including the creation of a functional requirements document and domain model. In order to capture the specificity of Punk culture, a preliminary genre vocabulary has been also been developed. The functional requirements document, domain model, and genre vocabulary will be published on a wiki for community discussion and feedback. The remaining phases of development, including the creation of a description set profile and usage guidelines, will be initiated following our review of community interest and comments.

The ultimate goal of this DCAP is to reach the Punk community and achieve broad adoption. The outcome of our work would aid in the effective acquisition and dissemination of Punk materials, or their metadata, in a variety of settings. Our project will also be useful to other niche communities documenting their cultural contributions because it provides a model that incorporates community outreach with traditional metadata development to lend more credibility and visibility to the end result.

AUTHOR INDEX

TDL.ORG
TEXAS DIGITAL LIBRARY

DC 2014 AUSTIN TEXAS