2016 Proceedings of the International Conference on Dublin Core and Metadata Applications

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

Copenhagen, Denmark
13-16 October 2016

ISSN: 1939-1366 (Online)

METADATA SUMMIT

WORKSHOPS

DC-1, Dublin, Ohio USA: 1-3 March 1995
DC-2, Warwick, UK: 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, U.S.A.: 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, USA, 8-11 October 2014
DC-2015, São Paulo, Brazil: 1-4 September 2015
DC-2016, Copenhagen, Denmark: 13-16 October 2016

ORGANIZING COMMITTEE

Danish DC-2016 Committee

Leif Andresen, Advisor to the Director, The Royal Library. National Library of Denmark
Preben Aagaard Nielsen, Danish Agency for Culture, Denmark
Susanne Thorborg, Danish Bibliographic Centre, Denmark

Technical Program Chairs

Valentine Charles, Europeana Foundation, Netherlands
Lars G. Svensson, Deutsche Nationalbibliothek, Germany

Professional Program Chairs

Thomas Baker, Dublin Core Metadata Initiative (DCMI), Germany
Michael D. Crandall, University of Washington, United States
Stuart A. Sutton, Dublin Core Metadata Initiative (DCMI), United States

Technical Program Committee

Leif Andresen, Advisor to the Director, The Royal Library. National Library of Denmark
Thomas Baker, Dublin Core Metadata Initiative (DCMI), Germany
Ana Alice Baptista, Universidade do Minho, Portugal
Joseph A Busch, Taxonomy Strategies, United States
Eric Childress, OCLC Research, United States
Michael D. Crandall, University of Washington, United States
Gordon Dunsire, Independent Consultant, United Kingdom
Kai Eckert, Hochschule der Medien (Stuttgart Media University), Germany
Kevin Ford, Art Institute of Chicago, United States
Christophe Guéret, BBC Wales
Corey A. Harper, New York University
Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Netherlands
Johannes Keizer, Food and Agriculture Organization of the United Nations (FAO), Italy
Wouter Klapwijk, Stellenbosch University, South Africa
Mariana Curado Malta, CEISE/ISCAP - Polytechnic of Oporto, Portugal
Marcia A. Mardis, Florida State University, United States
Filiberto Felipe Martinez-Arellano, National Autonomus University of Mexico, Mexico
Akira Maeda, Ritsumeikan University, Japan
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, United States
Steven J. Miller, University of Wisconsin-Milwaukee, United States
Jin-Cheon Na, Nanyang Technological University, Singapore
Liddy Nevile, Retired, Australia
Annelies van Nispen, Eye Film Institute, Netherlands
Johan Oomen, Netherlands Institute for Sound and Vision, Netherlands
Oknam Park, Sangmyung University, Republic of Korea, Korea, Republic Of
Cristina Pattuelli, Pratt Institute, United States
Susanna Peruginelli, Susanna Peruginelli Library consultancy, Italy
Vivien Petras, Humboldt-Universität zu Berlin, Germany
Magnus Pfeffer, Stuttgart Media University, Germany
Serhiy Polyakov, Weill Cornell Medicine - Qatar, Qatar
Sarah Potvin, Texas A&M University Libraries, United States
Jian Qin, Syracuse University, United States
KS Raghavan, PES Institute of Technology, India
Stefanie Ruehle, SUB Goettingen, Germany
Johann Wanja Schaible, GESIS - Leibniz-Institute for the Social Sciences, Germany
Jodi Schneider, University of Pittsburgh, United States
Ryan Shaw, University of North Carolina at Chapel Hill, United States
Shigeo Sugimoto, University of Tsukuba, Japan
Stuart A. Sutton, Dublin Core Metadata Initiative (DCMI), United States

TABLE OF CONTENTS

SESSION 5
Metadata Profiles

SESSION 6
Data Sharing & Identifiers

POSTERS (Peer Reviewed)

POSTERS (Best Practice)

Application Profiles & Practicalities of Data Integration

Presentation

Japanese Metadata Standards "National Diet Library Dublin Core Metadata Description (DC-NDL)": Describing Japanese Metadata and Connecting Pieces of Data

Saho Yasumatsu
National Diet Library, Japan
s-yasuma@ndl.go.jp

Akiko Hashizume
National Diet Library, Japan
hasizume@ndl.go.jp

Julie Fukuyama
National Diet Library, Japan
ju-fukuy@ndl.go.jp

Keywords: The National Diet Library; National library; Japanese library; metadata; Linked Data

1. About the NDL

The National Diet Library (NDL) is the sole national library in Japan. The NDL acquires, preserves and provides Japanese publications which are the nation's cultural and intellectual assets. The acquisition of library materials is mostly based on the Legal Deposit System. The NDL compiles and provides various bibliographies of library materials. Most of the collections are searchable through the NDL-OPAC and NDL Search on the website. To facilitate effective data use by computer systems or applications, the NDL initiatives to promote Linked Data and provides metadata as Linked Data.

This presentation shows the National Diet Library Dublin Core Metadata Description (DC-NDL), which defines elements and rules for the NDL's metadata. The views and opinions expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the NDL or related organizations.

2. What is DC-NDL?

The DC-NDL is a descriptive metadata standard that is utilized primarily for converting catalog records of publications held by the NDL into metadata based on the Dublin Core Metadata Element Set (DCMES) and the DCMI Metadata Terms. The DC-NDL comprises three parts: NDL Metadata Terms, a list of metadata terms defined by the NDL; Application Profiles, which clarify the standard use of the DC-NDL in the NDL systems; and RDF Schema, which is a Resource Description Format (RDF) version of the NDL Metadata Terms.

In this presentation, the authors focus mainly on the Application Profiles. The DC-NDL is structured as shown in Fig. 1. It is a triple-layer data model, comprising Administrative Information, Bibliographic Information, and Item Information. This presentation contains an explanation of Bibliographic Information.

FIG. 1. Basic structure of the DC-NDL
*FIG. 1. does not show the full elements, but basic elements of the DC-NDL.

2.1. A Brief History

The DC-NDL was developed in order to enhance the interoperability of metadata among Japanese libraries and related institutions. It has been revised three times (in May 2007, June 2010 and December 2011) since its launch in March 2001, to add metadata terms and redefine their usages, reflecting revisions to the Dublin Core and the expansion of NDL services.

3. Features of the DC-NDL

The DC-NDL requires mechanisms that can support the NDL's own mission, such as collecting and preserving Japanese publications. In other words, the DC-NDL has to accommodate not only certain unique features of the Japanese language but NDL activities. At the same time, the DC-NDL needs to maintain the broad perspective of linking to the world's data.

The key functions of the DC-NDL are the follows: (1) Representing the *yomi* (pronunciation) of the Japanese language, (2) Connectivity with Linked Data, and (3) Compatibility with digitized materials.

3.1. Representing *yomi* (pronunciation): characteristics of the Japanese language

The Japanese language has three distinct types of characters: *hiragana* (cursive syllabary), *katakana* (angular syllabary) and *kanji* (Chinese character).

いちご ／ イチゴ ／ 苺

FIG. 2. Example of *hiragana, katakana,* and *kanji* for "strawberry."
*All of these are pronounced "i-chi-go"

Additionally, there are cases in Japanese where pronunciation will vary depending upon meaning, so it is necessary to indicate the pronunciation using *hiragana* or *katakana*.

〝西瓜〟
●Pronunciation 1: su-i-ka
●Pronunciation 2: ni-shi-u-ri

FIG. 3. Example of a word pronounced differently from its character representation
*Watermelon in Japanese could be pronounced either "su-i-ka" or "ni-shi-u-ri"

Furthermore, the Japanese language generally does not have spaces inserted between words. This makes it very difficult to parse correctly when the entire sentence is written phonetically in *hiragana* or *katakana*.

ココデハキモノヲヌイデクダサイ
●Pronunciation 1: ココ デ ハ キモノ ヲ ヌイデ クダサイ
"Koko-de-wa-kimono (means clothes)-wo-nuide-kudasai"
●Pronunciation 2: ココ デ ハキモノ ヲ ヌイデ クダサイ
"Koko-de-hakimono (means shoes)-wo-nuide-kudasai"

FIG. 4. Example of a sentence with two different meaning, "clothes" or "shoes," depending on how the sentence is parsed.
*Pronunciation 1 means "Please take off your clothes here," Pronunciation 2 means "Please take off your shoes here."

The DC-NDL has been designed to serve as a national standard, and therefore defines metadata terms which can represent these characteristics of the Japanese language. Taking the Title field as an example, there are three salient features, as described below.

First, to describe the *yomi* (pronunciation) or transliteration of a particular title, the DC-NDL defines original properties such as a Transcription [dcndl:transcription], a Title Transcription [dcndl:titleTranscription], and an Alternative Transcription [dcndl: alternativeTranscription].

Second, the value stored in the Transcription [dcndl:transcription] property is the sentence with blanks between words, which is called *wakachi-gaki* in Japanese.

Third, the DC-NDL title representation and its pronunciation can be described as a set. Rather than using dcterms:title with restrictions on the literal range, dc:title is used without restrictions in its range. Describing a representation in a set with its *yomi* allows metadata to show semantic structures or logical relations of original catalog records, so that computers can handle the semantic structures and process them.

[Extract data from Bibliographic Information]

```
<dc:title>
  <rdf:Description>
    <rdf:value>謎の七支刀</rdf:value>
    <dcndl:transcription>ナゾ ノ シチシトウ</dcndl:transcription>
  </rdf:Description>
</dc:title>
```

FIG. 5. Example of a set comprising the representation and pronunciation of a title.

3.2. Connectivity with Linked Data

In the DC-NDL, Uniform Resource Identifier (URI) values can be used to link with other data. The following is an explanation of how to link an element to other data using a URI to link to creator and subject.

The DC-NDL creator property is intended to be used with URIs from Web NDL Authorities. Web NDL Authorities are a web service created by the authority data and maintained by the NDL. Web NDL Authorities include links to the Virtual International Authority File (VIAF), so that the authority data of the NDL can link to other major name authority files around the world through the VIAF.

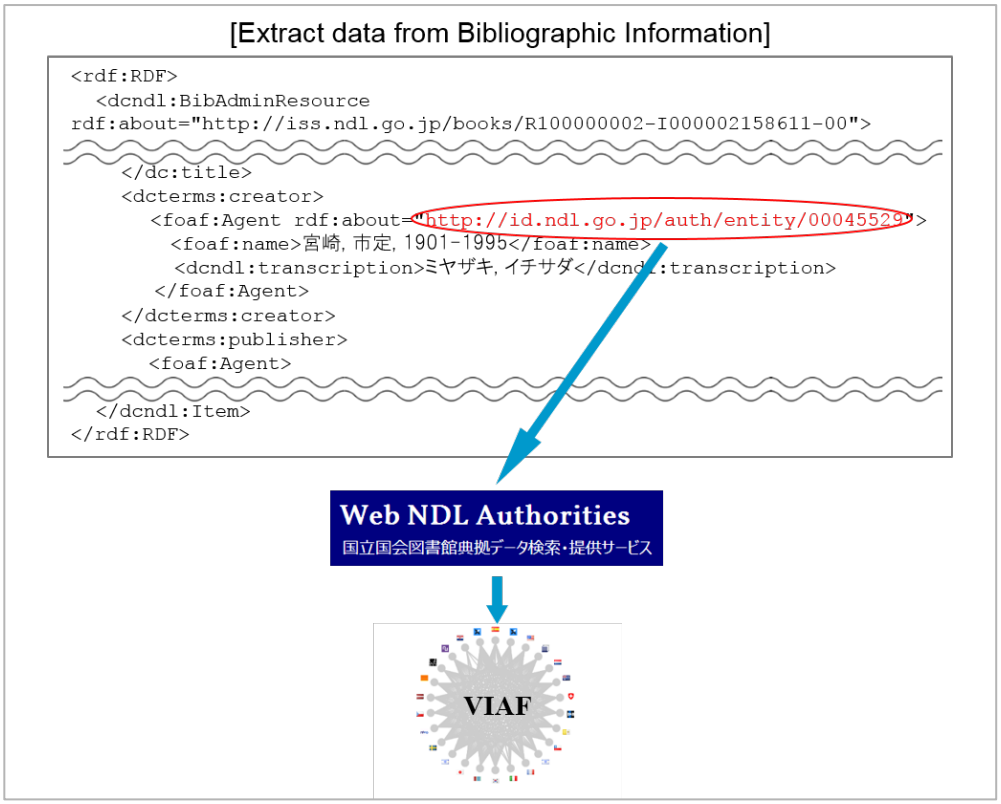


FIG. 6. Example of link to "dcterms:creator"

The DC-NDL subject property is intended to use URI values from classification systems adopted by the NDL, such as the Nippon Decimal Classification (NDC) 9th edition, the National Diet Library Classification (NDLC) and the Dewey Decimal Classification (DDC). By the way, the NDC is a Japanese standard classification, and the NDLC is developed by the NDL.

[Extract data from Bibliographic Information]

●Example 1: The NDC 9th edition

```
<dcterms:subject rdf:resource="http://id.ndl.go.jp/class/ndc9/210.02"/>
```

●Example 2: The NDLC

```
<dcterms:subject rdf:resource="http://id.ndl.go.jp/class/ndlc/GB38"/>
```

FIG. 7. Example of using URIs in "dcterms:subject"

In addition, the subject property can use URIs from the National Diet Library List of Subject Headings (NDLSH). NDLSH, which is compiled and maintained by the NDL, is a controlled subject vocabulary for accessing information resources from a subject. Most NDLSH subject headings link to the Library of Congress Subject Headings (LCSH) as a part of the Web NDL Authorities. So that, via LCSH the NDLSH links to subject headings of several national libraries in the world.

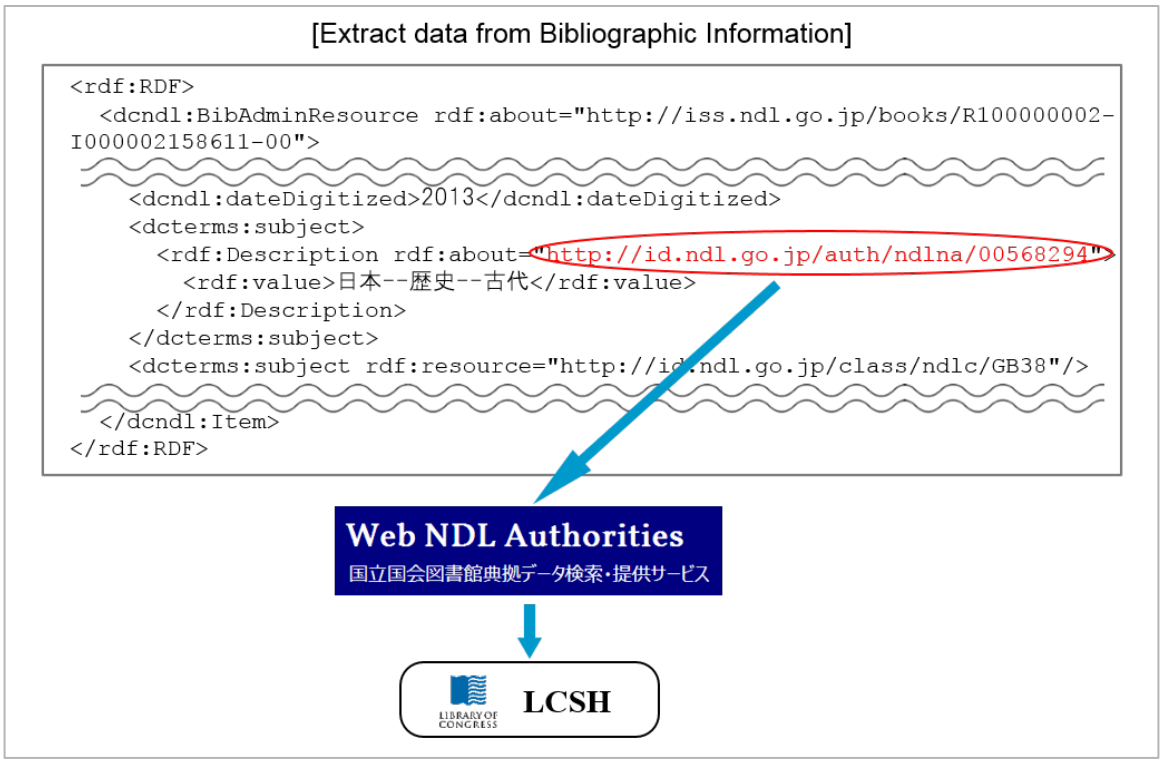


FIG. 8. Example of link to "dcterms:subject"

3.3. Compatibility with digitized materials

The NDL in making every effort to digitize its holding. To help describe those digitized contents adequately, the DC-NDL also defines original metadata terms.

[Extract data from Bibliographic Information]

Example of the original publisher of the material:

```
<dcterms:publisher>
  <foaf:Agent>
    <foaf:name>中央公論社</foaf:name>
    <dcndl:location>東京</dcndl:location>
  </foaf:Agent>
</dcterms:publisher>
```

●Example 1: Description of the publisher of the digitized version

```
<dcndl:digitizedPublisher>国立国会図書館</dcndl:digitizedPublisher>
```

●Example 2: Date digitized

```
<dcndl:dateDigitized>2013</dcndl:dateDigitized>
```

FIG. 9. Examples of terms for digitized content

4. Using the DC-NDL: NDL Search

4.1. What is NDL Search?

Finally, here is an example of implementing DC-NDL for use with NDL Search.

NDL Search is an integrated information search service that serves as a gateway to the rich repository of knowledge contained in the NDL, public libraries, academic libraries, archives, museums, and academic research institutions in Japan. It officially opened to the public on January 2012 and can search about 85 million metadata records as of March 2016. Data sources for the NDL Search include: NDL-OPAC, Japanese Periodicals Index, National Diet Library Digital Collections, digital archives provided by public and academic libraries in Japan, etc.

4.2. The API and its terms of use

NDL Search provides bibliographic data in RDF/XML of books, journals, articles, newspapers, and digitized material, audio files, web pages, and other digital content. This data includes title, author, publisher, subject matter, classifications, ISBN, ISSN, National Bibliography No., NDLJP (the NDL Digital Collection), URLs of webpages which show digitized content (http://dl.ndl.go.jp/...), and information related to copyright protection. The NDL Search accesses metadata in two encoding formats, DC-NDL (RDF) and DC-NDL (Simple), which are available from the user interface as well as via the API. The API can be used with the following protocols; SRU, SRW, OpenSearch, Z39.50 and OAI-PMH. The amount of information each encoding format provides is varied.

As of July 15, 2016, free use of metadata is predicated on the assumption that the use is non-commercial. People who wish to use the API of the NDL Search for a commercial activity are requested to apply for a license. Also, people who require continuous access to the API for a non-profit activity are requested to apply for a license.

5. The DC-NDL in the future: for connecting more data

The following are issues for future research on the DC-NDL:

The DC-NDL creator and subject fields link to other data, such as LCSH. To make the DC-NDL more linkable, we need to implement a means to add links. Thus, we must examine which elements are appropriate for linking with other data and which URIs are suitable for including in our data. Moreover, we need to identify data that can be linked with ours.

References

DC-NDL. Retrieved July 15, 2016, http://www.ndl.go.jp/en/aboutus/standards/meta.html

Web NDL Authorities. Retrieved July 15, 2016, from http://id.ndl.go.jp/auth/ndla

VIAF. Retrieved July 15, 2016, from http://viaf.org/

NDL Search. Retrieved July 15, 2016, from http://iss.ndl.go.jp/?locale=en

Presentation

Ontology Assessment and Extension: A Case Study on LD4L and BIBFRAME

Steven Folsom
Harvard Library, U.S.A.
steven_folsom@harvard.edu

Jason Kovari
Cornell University, U.S.A.
jak473@cornell.edu

Abstract

Representatives of the Andrew W. Mellon funded Linked Data for Libraries - Labs and Linked Data for Production teams will discuss their assessment strategy and alignment progress between the BIBFRAME and LD4L ontologies, including semantic patterns and ontology reuse. Further, the talk will discuss the ontology extension work underway within the LD4P program, focusing on those directed by Cornell and Harvard Universities.

Description

During the 2014-2016 Andrew W. Mellon funded Linked Data for Libraries (LD4L) project, the LD4L team created an ontology to express library resources as linked data. While existing ontologies addressed some necessary semantics and were thus reused, the LD4L team did not believe that any one of the existing bibliographic-focused ontologies matched overall semantics and patterns the team needed to express. A primary focus of this investigation included the first version of BIBFRAME; development on the LD4L ontology happened concurrently to that of BIBFRAME v2 (BF2) with the intent that the team would assess BF2 upon its release.

Since April 2016, representatives of the Mellon funded Linked Data for Libraries - Labs (https://www.ld4l.org/ld4l-labs/) and Linked Data for Production (https://www.ld4l.org/ld4p/) teams have been assessing BF2 in an effort to align semantics between BF2 and the LD4L ontology. As part of this alignment investigation, the team reviewed principles followed by the BIBFRAME architects as well as those important in the wider linked data world. Further, the team has investigated ontology reuse and external ontology subclassing assertions within BF2 as well as in depth specifics around semantic patterns. The ultimate goal of this alignment is to influence improvements to BIBFRAME, prune the LD4L ontology and identify a target ontology for LD4L/LD4P tooling to support native linked data production and conversion efforts.

Another ontology-related component of the LD4P and LD4L Labs projects include ontology extensions for focused areas of description; more information on this work can be found on the LD4P wiki site (https://wiki.duraspace.org/x/VQJxB). This extension work will recommend implementation patterns for describing resources in particular domains with greater specificity than designed to be provided by the core BIBFRAME framework. Representatives of the Cornell and Harvard Universities directed extensions will detail this process as well as goals for the extensions for the areas of Rare, Cartographic/Geospatial, Moving Image materials.

Presentation

Save the Children Resource Libraries: Aligning Internal Technical Resource Libraries with a Public Distribution Website

Joseph A. Busch
Taxonomy Strategies, U.S.A.
jbusch@taxonomystrategies.com

Branka Kosovac
Taxonomy Strategies, U.S.A.
bkosovac@taxonomystrategies.com

Katie Conrad
Save the Children, Sweden
Katie.Konrad@rb.se

Martin Svensson
Save the Children, Sweden
Martin.Svensson@rb.se

Abstract

Save the Children (STC) is an international NGO that promotes children's rights, provides relief and helps support children across the globe. With international headquarters in London, STC has 30 national members and supports local partners operating in over 100 countries worldwide. STC International maintains technical infrastructures that are available to members and local partners including SharePoint, Drupal and other information management applications. An effort to specify and implement a common resource library for curating and sharing internal technical resources has been underway since November 2015. This has included an inventory of existing (but heterogeneous) resource libraries on Save the Children's work in the thematic area of Health and Nutrition, and agreement on a common metadata specification and some controlled vocabularies to be used going forward. This internal technical resource library has been aligned with Save the Children's Resource Centre (http://resourcecentre.savethechildren.se/), a public web-accessible library that hosts comprehensive, reliable and up-to-date information on Save the Children's work in the thematic areas of Child Protection, Child Rights Governance and Child Poverty. The goal is to make it easy for content curators to identify items in the internal technical resource library, and to publish them to the public Resource Centre with a minimum transformation of metadata required. This presentation will discuss how this project has reached consensus on how to accommodate and balance internal research and external communication requirements by developing a light-weight application profile.

Bios

Joseph Busch is the Founder and Principal Consultant of Taxonomy Strategies. Taxonomy Strategies guides global companies, government agencies, and NGO's such as Kraft Foods, the Center for Medicare and Medicaid Services, and the Robert Wood Johnson Foundation in developing metadata frameworks and taxonomy strategies to help information achieve its highest value. Before founding Taxonomy Strategies, Mr. Busch held management positions at Interwoven, Metacode Technologies, the Getty Information Institute, PriceWaterhouse and Hampshire College. He is a Past President of the Association for Information Science and Technology, and a member of the Dublin Core Metadata Initiative Executive Committee.

Branka Kosovac is a Taxonomy Strategies associate and Principal of dotWit Consulting. She develops and implements complex taxonomies in a variety of business and technical contexts for Fortune 100 companies, international organizations, government agencies and mid-size enterprises across North America and the European Union such as Microsoft Corporation, Canadian National Research Council, Ford Foundation, and United Nations Development Program. Branka has mentored numerous taxonomy consultants, developed methodologies,

established the taxonomy practice for larger consulting companies, and taught as an Adjunct Professor at the University of British Columbia.

Katie Konrad is the Cataloging Librarian for Save the Children's Resource Centre, a digital library of child rights materials. As resident metadata enthusiast, Katie has been a key advocate for the metadata and taxonomic integrity of the Resource Centre. She holds a master's degree in Digital Library and Information Sciences, as well as a master's degree in International Relations and Political Science. Before joining Save the Children, she was Lead Consultant for the Uppsala Conflict Data Programme, as well as Intern Librarian for IDEA, Stockholm.

Martin Svensson is the Resource Centre Manager and has been working on the overall strategy and development of the website for the past three years. Focus has been on making a more scalable structure, improving data quality and taxonomies, and innovating the UX and UI. This has led to better Knowledge Management inside Save the Children, as well as increased number of external users. Before Save the Children, Martin worked as Regional Manager for Young Enterprise/Junior Achievement and as Usability Designer at Nokia Home Communications.

Presentation

Metadata for 3D Geological Models: Definition and Implementation

Etienne Taffoureau
Bureau de Recherches Géologiques et Minières, France
e.taffoureau@brgm.fr

Christelle Loiselet
Bureau de Recherches Géologiques et Minières, France
c.loiselet@brgm.fr

Abstract

BRGM (the French geological survey) is France's reference public institution for Earth Science applications which works on management and delivering geosciences data to be used for helps to decision-making for spatial planning, mineral prospecting, groundwater prospecting and protection, pollution control, natural risk prevention and the characterization of local areas.

Some of this data are produced from 3D geological modeling which is now a classical tool to better constrain geometries of complex geological systems and provide a continuous description of the subsurface out of sparse and indirect data. In order to store and deliver geological model production at BRGM, we developed a programming interface distinguishing the storage of the model from the representation of the model: models are stored using native format of the tool used to generate with (software project files). This choice guarantees that there is neither loss of data nor loss of precision. Then, model discretization (e.g. meshes) are generated on demand, depending on representation purposes (1, 2 or 3D gridding). Geological organization works on geomodel management and their representation for delivering and disseminating 3D geological information.

Therefore, it needs to reference and archive geo models and / or representation to access and to deliver information related to.

We propose to define a metadata profile compliant with INSPIRE[1] Directive to describe 3D geological models and their representation. The profile is implemented using the ISO 19115/19139 standard (used for geographic data) (1) to allow web application to edit and to manage data with GeoSource/GeoNetwork application; (2) to ensure interoperability in the delivery. 3D geomodel metadata are indexed by a search engine and displayed in a geoscientific portal such as Infoterre (http://infoterre.brgm.fr/viewer). Our approach allows calling the programming interface which queries 3D geological model and retrieves all the topological information from the model to be represented and stored or visualized by using OGC standards.

Our research work is linked to international initiatives (such as (i) OGC[2]; IUGS / CGI[3] for standard and (ii) One Geology[4] and EPOS[5] projects to test implementation) to define an interoperable model and to ensure common metadata for geological models.

[1] http://inspire.ec.europa.eu/
[2] http://www.opengeospatial.org/
[3] http://www.cgi-iugs.org/
[4] http://www.onegeology.org/
[5] https://www.epos-ip.org/

Presentation

Metadata on Biodiversity: Definition and Implementation

Etienne Taffoureau
BRGM, Bureau de Recherche Géologique et Minière, France
e.taffoureau@brgm.fr

A. Cohen Nabeiro
FRB, Fondation pour la Recherche sur la Biodiversité, France

J. Touroult
MNHN, Muséum National d'Histoire Naturelle, France

Abstract

SINP (Information system on nature and landscape) and ECOSCOPE (Observation for research on biodiversity data hub) are two distinct scientific infrastructures on biodiversity relying on different data sources and producers. Their main objective is to document and share information on biodiversity in France. INPN (https://inpn.mnhn.fr) is the reference information system for data related to nature. It manages and disseminates the reference data of the "geodiversity and biodiversity" part of the SINP, and deliver the metadata and data to GBIF (Global Biodiversity Information Facility). For SINP and Ecoscope projects, working groups composed of scientific organisations have defined two compliant metadata profiles, also compliant with INSPIRE Directive, to describe data on this thematic. These profiles are implemented using existing metadata standards: ISO 19115/19139 (for geographic metadata) for SINP and EML (Ecological Metadata Language) and ISO 19115/19139 for ECOSCOPE. A mapping has also been processed between the two profiles, as well as several thesaurus for keywords and a classification system for taxonomic identification are used, so as to ensure interoperability between systems. The profiles are implemented in web applications for editing and managing data (GeoSource/GeoNetwork for SINP and an ad hoc application for ECOSCOPE). These applications allow the harvesting of metadata using OGC/CSW (Catalog Service for the Web) standard.

Next steps will permit to increase metadata visibility through the automatization of web-services.

Maintaining Vocabularies: Define, Release & Curate

Presentation

Data-Driven Development of the Dewey Decimal Classification

Rebecca Green
OCLC Online Computer Library
greenre@oclc.org

Description

Changes involved in maintaining the Dewey Decimal Classification (DDC), a general classification system, have derived in the past from many distinct sources. These include (but are not limited to) questions/ideas/complaints from end users, classifiers, translators, or members of the Decimal Classification Editorial Policy Committee (EPC); mappings of other knowledge organization systems to the DDC; and personal awareness of events, emerging issues, and trends. On the one hand, these phenomena may bring to light ambiguity or redundancy in the current system. On the other hand, they may bring to the attention of the editorial team new topics needing provision within the system.

Without disregarding these sources, the DDC editorial team is also considering data-driven methods of (1) identifying existing areas of the DDC warranting further development or (2) identifying topics with sufficient literary warrant to justify explicit inclusion in the DDC. The use of two sources of data is under investigation.

The first data source reflects the assignment of recently created Library of Congress Subject Headings (LCSHs) to resources described in WorldCat records (i.e., LCSHs added within the past 5 years). Identifiable sets of headings typically not mapped to the DDC (e.g., personal, family, and corporate names) are filtered out; headings are further restricted to those appearing in at least 10 WorldCat records. For these we gather the number of records to which they are assigned, corresponding holdings data, and any numbers from the current full edition of the DDC that are assigned to the same records. Sorted by number of records or holdings, such a headings list helps prioritize development of the DDC by topic. Further massaging of the data in conjunction with the DDC's expressive notation isolates areas of the classification most associated with emerging topics and thereby helps prioritize development by area of the system.

The second data source reflects the assignment of numbers from the current full edition of Dewey to WorldCat records. For each DDC number, we compute a value that favors these conditions:

- The DDC number and built numbers for which the original number is the base number are assigned relatively more often than other DDC numbers.
- The DDC number has been assigned relatively more often than its subordinate numbers.
- The DDC number has been assigned relatively more often than built numbers (including standard subdivisions) for which it is the base number.

Sorting the list of DDC numbers by the computed value helps identify areas within the schedule which are receiving extensive use, but are not well developed.

The topics and schedule areas identified through these means require investigation to ascertain if they are viable candidates for further development. Preliminary work with these data sources reveals that the strategies hold promise.

The Global Agricultural Concept Scheme and Agrisemantics

Thomas Baker
Independent FAO consultant, Bonn, Germany
tom@tombaker.org

Caterina Caracciolo,
Food and Agriculture Organization of the UN,
Rome, Italy
caterina.caracciolo@fao.org

Anton Doroszenko,
CAB International,
Wallingford, UK
a.doroszenko@cabi.org

Lori Finch,
National Agricultural Library, USDA, USA
lori.finch@ars.usda.gov

Osma Suominen,
National Library of Finland,
Helsinki, Finland
osma.suominen@helsinki.fi

Sujata Suri
National Agricultural Library, USDA, USA
sujata.suri@ars.usda.gov

Abstract

Key concepts from three thesauri about agriculture and nutrition—AGROVOC, CAB Thesaurus, and NAL Thesaurus—have been merged into a Global Agricultural Concept Scheme (GACS). The respective partner organizations—Food and Agriculture Organization of the UN (FAO), CAB International (CABI), and the USDA National Agricultural Library (NAL)—undertook this initiative in 2013 with the goal of facilitating search across databases, improving the semantic reach of their databases by supporting queries that freely draw on terms from any mapped thesaurus, and achieving economies of scale from joint maintenance. The GACS beta release of May 2016 has 15,000 concepts and over 350,000 terms in 28 languages.

The creation of GACS began by mapping three sets of 10,000 frequently used concepts from the three thesauri to each other, pairwise. Mappings were vetted by experts; vetted mappings were algorithmically checked for awkward clusters, or "lumps"; and lumps were resolved through discussion on teleconferences and in meetings—for example, by drawing a line between "energy intake" (related to organisms) and "energy consumption" with the narrower "fuel consumption" (related to natural resources). Mappings were manually corrected, and GACS was iteratively re-generated, until the set of concepts was considered stable enough for publication as GACS Beta.

Some inevitable results of this process of aggregation, such as overlapping labels, have already been fixed. Other issues, such as concepts with multiple hierarchical relations ("polyhierarchy"), have yet to be tackled. The working group has revived a classification scheme, developed jointly in the 1990s, to tag concepts by thematic group. Concepts are being typed as chemical, geographical, organisms, products, or topics. Alongside generic thesaurus relations to broader, narrower, and related concepts, organisms will be related to relevant products.

GACS is seen as a first step for Agrisemantics, an emerging community network of semantic assets relevant to agriculture and food security. Within Agrisemantics, the general-purpose, search-oriented concepts of GACS are intended to serve as a hub for concepts defined, with more precision, in a multitude of ontologies modeled for specific domains. Ontologies, in turn, are intended to provide global identity to concepts used in a vast diversity of quantitative datasets, such as sensor readings and crop yields, defined for a multitude of software applications and serialization formats.

Such semantic authority control of data elements could support, for example, an analysis of the yield gap in sub-Saharan Africa. A wheat data element, labeled 'GW' in a phenotype dataset, could be mapped to the concept 'grain weight' as defined and globally identified in the CGIAR Crop Ontology. In turn, the Crop Ontology concept could be mapped to the broader concept 'Grain' in GACS. Searches could return not only datasets about grain weight, but references to published papers where the weight of the grain was studied.

Agrisemantics aims at improving the discoverability and semantic interoperability of agricultural information and data for the benefit of researchers, policy-makers, and farmers with the ultimate goal of enabling innovative responses to the challenges of food security under conditions of climate change. Achieving these goals will require innovation in processes for the cooperative maintenance of linked semantic assets in the modern Web environment.

Presentation

The Dutch Art & Architecture Thesaurus® Put into Practice: The Example of Anet, Antwerp

Karen Andree
University of Antwerp, Belgium
karen.andree@uantwerpen.be

Reem Weda
RKD – Netherlands Institute for Art History, The Netherlands
weda@rkd.nl

Abstract

Anet is a network of scientific libraries located in Antwerp, Belgium. Among the connected institutions are research, higher education and museum libraries. They share common software (Brocade, developed at the Antwerp University since 1998) and cataloging practices. In 2014 was decided to adopt a new subject heading system for cataloging the library collections with an art or heritage scope. The Art & Architecture Thesaurus® (maintained by the Getty Research Institute) was eventually selected, under the express condition that it can be used in a flexible way by the libraries. This includes, if needed, the usage of non-AAT compatible subject terms.

AAT was chosen because of software compatibility, extensiveness of its content and multilingualism. The thesaurus is being fully translated into Spanish, Dutch, German, Chinese, and partly in other languages, such as Italian and French.

The local subject heading systems (terminologies) were converted to the new authority environment (Anet-AAT). Automatic mapping via tools was considered. However, manual mapping was performed because of the different application as subject heading system and the opportunity to acquaint the librarians with AAT.

Future challenges for the Anet-AAT vocabulary consist of staying updated with changes that occur in the 'Mother AAT' (Getty Vocabularies) and adding to its content to create more library specific subjects - AAT is presently quite focused on the description of (museum) objects.- But, since Anet is using AAT, it's been noticed that the content is quite well suited for indexing the special libraries. Nevertheless, the usage by the network did bring to light issues in the structure of AAT, particularly some concerning the Dutch translation. The necessity to address these issues has resulted in regular contacts between Anet and the RKD-Netherlands Institute for Art History that manages the Dutch translation of the AAT.

The AAT has a long history of development. The original AAT by Getty already started in the late seventies in the United States. The RKD manages the Dutch AAT, or 'AAT-Ned' since the mid-nineties. Work on the expansion and improvement of the Thesaurus is an ongoing process. Being a 'living terminology', this has impact on the usage as a standard by others. The publication of the Getty Vocabularies as Linked Open Data only made this more apparent. Together with user communities such as Anet the RKD tries to adapt the content of the AAT for the better. Particularly concerning the Dutch translation, but it also tackles other issues concerning scope notes (definitions), or hierarchical relations that are not compatible with the views of the Dutch-speaking heritage community. Because of the scope and size of the content of AAT The RKD cannot discover all the issues by itself, and needs the input of users from the heritage community to improve the system. In this manner, Anet contributes to the improvement of the Dutch translation as well as the 'mother- AAT'. The adaptation of AAT by Anet has proven to be a promising showcase for the potential of this 'museum thesaurus' as a subject heading system for libraries as well.

Keywords: metadata; thesaurus; subject headings; libraries; art libraries; Art & Architecture Thesaurus®; Brocade Library Services; authority control"

Presentation

A Study on the Best Practice for Constructing a Cross-lingual Ontology

Yi-Yun Cheng
Department of Library and Information Science, National Taiwan University, Taiwan
r02126002@ntu.edu.tw

Hsueh-Hua Chen
Department of Library and Information Science, National Taiwan University, Taiwan
sherry@ntu.edu.tw

Abstract

Ontologies, as the fundamental building blocks for the Semantic Web, are the highest-level classification scheme in the family of Knowledge Organization Systems (KOS). With the emergence of big data, ontologies are one of the keys to unraveling the information explosion problems. Under the big data situation, many language cultures are in a pressing need to construct ontologies. Cross-lingual ontology research has thus become a pivotal concern in this global age. Researchers worldwide try to be interoperable with ontologies written not only in English, but also in other languages. Yet, constructing a cross-lingual ontology can be difficult, and a detailed mapping method is often hard to find. The purpose of this study is to establish a feasible practice on building cross-lingual ontologies. The study will focus on the construction of an English-Chinese ontology from an existing source ontology and a KOS source. This study will also address the synonymy and polysemy problems of the target language (Traditional Chinese).

By adopting a three-phase research design, this study begins with the pretest of our mapping practice on a small ontology—W3C's Semantic Sensor Network Ontology (SSN ontology). This phase is to ensure our SPARQL code to parse all the classes in SSN ontology is feasible. In phase two, we try to map our source ontology—the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies—with the KOS term-lists from National Academy of Educational Research (NAER) in Taiwan. In phase three, we model the mapped English/Chinese ontology in Protégé software to explore the prospect of this method.

The results in phase one shows that our SPARQL code can automatically helped us retrieve all 117 classes in SSN ontology into plain text format in a click, suggesting that our practice is a workable one. In phase two and three, a cross-lingual ontology between English and Traditional Chinese is constructed through the implementation of Protégé. The mapping results between the 3,770 SWEET ontologies classes (in English) and the NAER term-lists (in Traditional Chinese) reveal an accuracy of 80.66% on the exact-match terms, while the Chinese synonyms and related terms expressed by SKOS labels are all proven searchable in our primary evaluation. These promising results demonstrate the feasibility of the practice proposed by this study, and further suggest that such approach is suitable to be adopted by future researchers to model their cross-lingual ontologies.

Keywords: ontologies; cross-lingual ontologies; SWEET ontologies

Presentation

Remixing Archival Metadata Project (RAMP) 2.0: Recent Developments and Analysis of Wikipedia Referrals

Mairelys Lemus-Rojas
Indiana University-Purdue University
Indianapolis, USA
mlemusrojas@gmail.com

Abstract

The RAMP (Remixing Archival Metadata Project) tool, developed at the University of Miami Libraries, emerged as a way of facilitating the contribution of library data to the English Wikipedia in alignment with the increasing interest in sharing and exposing distinctive library collections in the online encyclopedia. RAMP is an open source web-based editor that extracts biographical information from EAD (Encoded Archival Description) finding aids using the EAC-CPF (Encoded Archival Context-Corporate Bodies, Persons, Families) format. It also allows for the integration of additional data from other sources like WorldCat Identities and VIAF (Virtual International Authority File) and transforms all the information into wiki markup for publication to the English Wikipedia through its API.

In 2014, a pilot project was conducted using the Cuban Heritage Collection (CHC) Theater Collections. Google Analytics was used to track usage and referrals from Wikipedia to the University of Miami finding aids website, and a noticeable increase in traffic was seen. A report of the results of the pilot project was presented at the Fonds & Bonds DCMI Preconference in 2014. Later, the tool was further developed and has been used to contribute additional collections to Wikipedia. RAMP 2.0 was recently released, and a number of issues identified during a round of usability testing conducted at the library were addressed. This presentation will cover an analysis of referrals from all Wikipedia pages created using the tool. It will also feature a demo of the tool, and will highlight some of the recent developments, which include a major overhaul of the interface, more secure Wikipedia log in, easy upload capabilities, and an effective and convenient installation process. With this recent development, we are providing the library community with a tool that is easy to use and install and that offers a convenient way to share data with other communities on a global scale.

Linked Data for Data Integration and Curation

Presentation

POSTDATA – Towards publishing European Poetry as Linked Open Data

Mariana Curado Malta
Polytechnic of Oporto, Portugal
UNED-LINHD, Madrid, Spain
mariana@iscap.ipp.pt

Elena Gonzalez-Blanco
UNED-LINHD, Madrid, Spain
egonzalezblanco@flog.uned.es

Abstract

POSTDATA is a 5 year European Research Council (ERC) Starting Grant Project that began in May 2016 and is hosted by the Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain. The context of the project is the corpora of European Poetry (EP), with a special focus on poetic materials from different languages and literary traditions. POSTDATA aims to offer a standardized model in the philological field and a metadata application profile (MAP) for EP in order to build a common classification of all these poetic materials. The information of Spanish, Italian and French repertoires will be published in the Linked Open Data (LOD) ecosystem. Later we expect to extend the model to include additional corpora.

There are a number of Web Based Information Systems in Europe with repertoires of poems available to human consumption but not in an appropriate condition to be accessible and reusable by the Semantic Web. These systems are not interoperable; they are in fact locked in their databases and proprietary software, not suitable to be linked in the Semantic Web.

A way to make this data interoperable is to develop a MAP in order to be able to publish this data available in the LOD ecosystem, and also to publish new data that will be created and modeled based on this MAP. To create a common data model for EP is not simple since the existent data models are based on conceptualizations and terminology belonging to their own poetical traditions and each tradition has developed an idiosyncratic analytical terminology in a different and independent way for years. The result of this uncoordinated evolution is a set of varied terminologies to explain analogous metrical phenomena through the different poetic systems whose correspondences have been hardly studied – see examples in González-Blanco & Rodríguez (2014a and b). This work has to be done by domain experts before the modeling actually starts. On the other hand, the development of a MAP is a complex task though it is imperative to follow a method for this development. The last years Curado Malta & Baptista (2012, 2013a, 2013b) have been studying the development of MAP's in a Design Science Research (DSR) methodological process in order to define a method for the development of MAPs (see Curado Malta (2014)). The output of this DSR process was a first version of a method for the development of Metadata Application Profiles (Me4MAP) (paper to be published). The DSR process is now in the validation phase of the Relevance Cycle to validate Me4MAP (for more information and detail on DSR see Hevner (2007)). The development of this MAP for poetry will follow the guidelines of Me4MAP and this development will be used to do the validation of Me4MAP.

The final goal of the POSTDATA project is: i) to be able to publish all the data locked in the WIS, in LOD, where any agent interested will be able to build applications over the data in order to serve final users; ii) to build a Web platform where: a) researchers, students and other final users interested in EP will be able to access poems (and their analyses) of all databases; b) researchers, students and other final users will be able to upload poems, the digitalized images of manuscripts, and fill in the information concerning the analysis of the poem, collaboratively contributing to a LOD dataset of poetry.

Aknowledgements

The work presented in this abstract has been developed thanks to the research projects funded by MINECO and led by Elena González-Blanco: Acción Europa Investiga EUIN2013-50630: Repertorio Digital de Poesía Europea (DIREPO) and FFI2014-57961-R. Laboratorio de Innovación en Humanidades Digitales: Edición Digital, Datos Enlazados y Entorno Virtual de Investigación para el trabajo en humanidades, and the Starting Grant ERC-2015-STG-679528 POSTDATA.

References

Curado Malta, M. (2014). Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web Semântica. University of Minho. Retrieved from http://hdl.handle.net/1822/30262

Curado Malta, M., & Baptista, A. A. (2012). State of the Art on Methodologies for the Development of a Metadata Application Profile. Metadata and Semantics Research, 343(July), 61–73. doi:10.1007/978-3-642-35233-1_6

Curado Malta, M., & Baptista, A. A. (2013a). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In M. Foulonneau & K. Eckert (Eds.), Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013 (pp. 90–103). Lisbon: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81

Curado Malta, M., & Baptista, A. A. (2013b). Me4DCAP V0. 1: A method for the development of Dublin Core Application Profiles. In Information Services and Use (Vol. 33, pp. 161–171). Dublin Core Metadata Initiative. doi:10.3233/ISU-130706

González-Blanco, E., & Rodríguez, J. L. (2014a). ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse. Journal of the Text Encoding Initiative, (8).

González-Blanco, E., & Seláf, L. (2014b), Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires. In Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world, eds. L. Soriano - M. Coderch - H. Rovira - G. Sabaté - X. Espluga. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang, 321-332.

Hevner, A. R. (2007). The three cycle view of design science research. Scandinavian Journal of Information Systems, 19(2), 87.

Presentation

Cognitive and Contextual Computing Laying A Global Data Foundation

Richard Wallis
Data Liberate, United Kingdom
richard.wallis@dataliberate.com

Abstract

A search of the current computing and technology zeitgeist will not have to look far before stumbling upon references to Cognitive Computing, Contextual Computing, Conversational Search, the Internet of Things, and other such buzz-words and phrases. The marketeers are having a great time coming up with futuristic visions supporting the view of computing becoming all pervasive and 'intelligent'. From IBM's Watson beating human quiz show contestants, to the arms race between the leading voice-controlled virtual assistants – Siri, Cortana, Google Now, Amazon Alexa.

All exciting and interesting, but what relevance has this for DCMI, metadata standards, and the resources we describe using them? In a word, "context". No matter how intelligent and human-like a computer is, it's capabilities are only as good as the information it has to work with. If that information is constrained by domain, industry specialised vocabularies, or a lack of references to external sources; it is unlikely the results will be generally useful.

In the DCMI community we have expertise in sharing information within our organisations and on the web. Dublin Core being one of the first widely adopted generic vocabularies. A path that Schema.org is following and in its breadth of adoption is now exceeding.

From his wide experience working with Google, OCLC, European and National Libraries, the banking industry and others, Richard will explore new initiatives and the processes being undertaken to prepare and widely share data in a generally consumable way on the web.

Schema.org has been a significant success. Used by over 12 million domains, on over a quarter of sampled pages. It is enabling a quiet revolution of preparing and sharing data to be harvested into search engine Knowledge Graphs. Knowledge Graphs that power Rich Snippets, Knowledge Panels, Answer Boxes, and other search engine enhancements. Whilst delivering on one revolution, it is helping to lay the foundations of another.

Building a global web of interconnected entities, for intelligent agents to navigate, these Knowledge Graphs fed by the information we are starting to share generically, are providing the context that will enable Cognitive, Contextual and associated technologies scale globally. Ushering in yet another new technology era.

Presentation

A Pilot Study on Linked Open Data in Cultural Heritage: A Use Case of the Taiwan Digital Archives Union Catalogue

Sophy Shu-Jiun Chen
Academia Sinica, Taiwan
sophy@sinica.edu.tw

Abstract

The Taiwan Digital Archives Union Catalogue (http://catalog.digitalarchives.tw/), with more than 5 million digitized objects described with Dublin Core-based metadata, comes from the Taiwan E-Learning and Digital Archives Program (TELDAP) which was built on a national scale over the past 15 years. Academia Sinica Center for Digital Cultures (ASCDC)(http://ascdc.sinica.edu.tw/en/) is now in charge of the sustainable operation. The presentation aims to report how we adopt Lined Open Data (LOD) approach to publish these structure data, in order to make metadata and the digitalized objects get connected with related resources in the world.

The Taiwan Digital Archives, similar to the Europeana, has collected digitized collections from more than 100 libraries, archives, museums, academic institutions, and government agencies, such as the National Central Library, Academia Historica and National Palace Museum. The collection includes books, newspapers, artworks, photos, specimen and sounds. Most of the metadata descriptions and contents are in Chinese and are Asian culture oriented. In the LOD initiative, 850 thousand records with Creative Commons licensing have been selected as experimental pilot since January 2016.

The presentation will report 72 collections across 16 categories such as biodiversity, photos, architecture, anthropology, rare books, Buddhist texts and paintings, discussing the LOD design methods, issues, outcomes of the preliminary results and lessons learned, which covers the data model, cleaning for data quality, reconciling, publishing and applications. In addition, the different ways of LOD applications will also be demonstrated including online exhibitions and the reuse in digital humanities researches

Presentation

A Survey of Metadata Use for Publishing Open Government Data in China

Li Yuan
Department of Information Management
School of Public Administration
Sichuan University, China
yuanli@scu.edu.cn

Wei Fan
Department of Information Management
School of Public Administration
Sichuan University, China
fanw@scu.edu.cn

Abstract

Open government data (OGD) is one important type of open data which grows fast all round the world. Many governments and organizations have already put their data online to the public. At the same time, linked data which conducted by W3C provides the publish mechanism and technical recommendation to explore the linkage of open data. Linked data promote the openness and availability of open data. Currently, 1,443 government related datasets are retrieved from datahub.io (2016-7-1).

From document to dataset, metadata still plays key functions for describing, locating and managing OGD. Although most OGD has some basic categories, tags and properties, more comprehensive metadata vocabularies need to further study. Utilizing metadata to achieve high quality, findable, machine readable and understandable OGD is the fundamental task for government chief data officers.

In recent 5 years, there are some remarkable development of OGD in China. National Bureau of Statistics of China has built national data portal for publishing monthly, quarterly and annual data, as well as the regional data and census data, which has nearly 8 million data. Beijing, Shanghai, Wuhan and Guiyang cities provide public data service. Zhejiang province integrates the public data category. In future, China national open government data portal will be established in 2018 from *Promote the development of Big Data Platform for Action (2015)*.

This presentation will report the state of metadata use of OGD in China. We investigate 8 typical cases of China OGD which includes three levels (nation, province and city). It contains three parts as following:

1. Analyze the actual usage of metadata elements for datasets and data entry in these selected OGD portals and point out some usage issues;
2. Discuss the adaption of existed metadata standards (DCAT, Schema.org, GILS and etc.) for China OGD and propose a metadata vocabulary for China OGD.
3. Comparatively analyze data share and application of OGD between US and China by metadata interoperability.

References

Tim berners Lee (2009). Linked Data: Design Issues. Retrieved, July 1, 2016, from https://www.w3.org/DesignIssues/LinkedData.html.

Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space: Theory and Technology. Volume 1. Morgan & Claypool Publishers (2011). http://linkeddatabook.com/.

Pierre-Yves Vandenbussche, Bernard Vatant. (2012). Metadata Recommendations For Linked Open Data Vocabularies. https://lov.okfn.org/Recommendations_Vocabulary_Design.pdf.

The State Council of The people's republic of China. Promote the development of Big Data Platform for Action. Retrieved, July 1, 2016, from http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

Metadata Reuse & Metadata Quality

The British National Bibliography: Who uses our Linked Data?

Corine Deliot
The British Library,
United Kingdom
corine.deliot@bl.uk

Neil Wilson
The British Library,
United Kingdom
neil.wilson@bl.uk

Luca Costabello
Fujitsu Ireland, Ireland
luca.costabello@ie.fujitsu.com

Pierre-Yves Vandenbussche
Fujitsu Ireland, Ireland
pierre-yves.vandenbussche@ie.fujitsu.com

Abstract

The British Library began publishing a Linked Open Data (LOD) version of the British National Bibliography (BNB) in 2011 as part of its open metadata strategy. Although organisational benefits have been gained, it has been challenging to identify how data is used and by whom. System logs capture basic information and anecdotal usage is received via user feedback, but a lack of analytics tools has made it difficult to gain an understanding of service usage to support sustained investment. This paper describes a project between the British Library and Fujitsu Ireland that examined the insights gained from the development and application of Linked Data analytics. The results suggest such analytics offer LOD publishers many benefits, the most important being an ability to accurately assess service impact and target limited resources more effectively. By doing so publishers can begin to manage LOD services as efficiently as their web counterparts and continue the realisation of Linked Data's potential.

Keywords: British Library; library; Linked Open Data; publication; usage analysis; analytics

1. Introduction

The British Library and Linked Open Data: The British Library is the national library of the United Kingdom with responsibility for distributing metadata describing its collections and recording UK publishing output in the British National Bibliography (BNB)[1]. Originally, these services were aimed at the library community and operated on a commercial basis. However, in 2010 the British Library adopted an open metadata strategy in response to Government calls for improved access to public sector data to promote transparency, economic growth and research. It was also believed that enabling the wider re-use of library data would increase its community value, improve access to information and culture and maintain the relevance of library services.

Simultaneously with the Library's open data initiative was a growing interest in Linked Data's potential for creating new information resources and reaching new users. Such opportunities were felt compelling enough to warrant practical experimentation. Despite a steep technical learning curve for library staff, a Linked Open Data (LOD) representation of the BNB was launched in 2011. The move proved influential among the library community in moving the Linked Data 'debate' from theory to practice (Alemu et al., 2012). Unlike other experimental services, the LOD BNB has continued to evolve with regular monthly updates, the inclusion of new links (e.g. to the International Standard Name Identifier (ISNI)[2]) and content (e.g. serials). The value of the Library's work was recognised by the dataset getting a five star openness rating on Data.gov.uk[3]

[1] http://bnb.data.bl.uk

[2] http://www.isni.org

[3] https://data.gov.uk/dataset/the-linked-open-british-national-bibliography

and being awarded Open Data Institute certification[4]. Such recognition supported the justification for continued work on Linked Data at a time when resources were under considerable pressure from the economic downturn. However, it has always been recognised that continued funding would inevitably depend on hard evidence of significant levels of regular and systematic usage to prove the service met community needs.

Assessing the value of Linked Data services: Despite a continuing interest in Linked Data and numerous related projects, RDF data continues to appeal to a more specialist audience than other, simpler open data formats (e.g. .CSV). The true value of Linked Data services has also been difficult to quantify with limited options available for Linked Data publishers keen to find out how their triples are used and which user groups are attracted to them. Due to the open access approach, evolving technologies and new usage patterns, it can be difficult to accurately gauge service impact. Similarly, attribution may be problematic when services are assembled from multiple sources. The situation can also be complicated by activity-based charging for Linked Data hosting platforms where it is vital to distinguish between innovative forms of legitimate use and abusive activity requiring preventative action.

Access logs have been generated since 2011 and indicate BNB LOD usage can vary significantly from a few hundred thousand to several million transactions per month for no clear reason. The limited availability of LOD analytics tools compared to web equivalents restricted the value of these logs and their interpretation has been a continuing challenge due to the time and effort required to extract useful information. The absence of reliable analytics has made it difficult to clearly identify and prioritise system developments with anecdotal feedback occasionally taking the place of quantitative information on usage patterns or client applications. From a resource management perspective, the requirement to justify all expenditure in a difficult economic environment makes such information increasingly important. Similarly, usage-based charging for hosting coupled with significant variations in usage patterns can make accurate capacity planning problematic. Even determining the impact of raw LOD dumps offered in parallel to the endpoint has proven difficult due to the unwillingness of users to register for access or offer feedback. Interestingly this situation is contrasted by the willingness of over 1500 global users to register for the Library's Z39.50[5] open library data service[6].

A collaborative investigation of Linked Data analytics: A key theme of the Library's open metadata strategy is collaboration to promote experimentation beyond the library domain. The Library is particularly interested in areas where partners can offer rare insights or technical expertise. An offer from Fujitsu Research to collaborate in the exploration of Linked Data analytics was therefore welcomed as an opportunity to both examine BNB LOD usage and to potentially develop tools of interest to the wider LOD community. The results of the collaboration together with their potential for assisting other publishers of Linked Open Data are described below.

2. Publication as Linked Open Data

This section offers background on the Library's LOD publication: its architecture, data model and challenges encountered. Full details can be found in a previous publication (Deliot, 2014).

[4] https://certificates.theodi.org/en/datasets/1063/certificate

[5] Z39.50 refers to ISO 23950 and ANSI/NISO Z39.50. It is an international standard client/server protocol developed by the library community and maintained by the Library of Congress for searching and retrieving records from remote bibliographic databases. http://www.loc.gov/z3950/agency/

[6] http://www.bl.uk/bibliographic/datafree.html#m21z3950

2.1. Challenges

When the BNB LOD project started in 2010 the world of library Linked Data was evolving rapidly with little consensus on many issues, e.g. re-use existing ontologies or create your own? (Hannemann and Kett, 2010). Some challenges arose from a requirement to work with converted legacy data since numerous changes in technology and standards over the 60+ years of BNB's existence necessitated careful normalisation processes. In addition, a transition from the flat data structure of the library domain MARC21[7] (MAchine Readable Cataloging) format to the open RDF entity-based model was problematic as, despite its name, MARC21 was not designed for machine actionability as currently understood. Assigning URIs to bibliographic entities originally represented as text strings involved compromises imposed by the available tools. Inevitably, some challenges also arose from the data modelling decisions made, e.g. imposing formal structure on transcribed text - a more complex process than treating it as a literal. Due to the steep learning curve, it was decided to concentrate on data modelling and conversion activities and use an externally-hosted SPARQL endpoint offered by Talis[8]. This practice continued with the later migration to the TSO OpenUp platform[9].

2.2. Data modelling

When the British Library decided to publish the BNB as Linked Open Data, there was little internal expertise in RDF or domain modelling. The Library therefore used Talis to train and mentor staff and assist development of data models for books[10] and serials[11]. The modelling process stepped back from MARC21 concepts to identify what such records expressed about "things in the world", whether concepts or material objects e.g. bibliographic resources, persons, organisations, etc. The intention was to model a defined part of the bibliographic domain accurately rather than just convert MARC21 to RDF and to focus on the main entities present in the data rather than attempt to replicate MARC21's complex structure and content.

The British Library data model has two main features. Firstly, in order to make the dataset useful beyond the library domain, resources such as books and serials are modelled in accordance with the popular understanding of their meaning rather than more abstract models (e.g. FRBR[12]). Secondly, publication is modelled as an event due to known future requirements to represent forthcoming publications and extend the model to cover further lifecycle events (e.g. acquisition, launch, etc.). To increase interoperability and minimise the overhead of maintaining an extensive British Library ontology, entities and relationships were described using existing RDF vocabularies and ontologies (e.g. Dublin Core, FOAF, etc.). New classes and properties were only defined and documented in the British Library RDF schema[13] where required for the data model. Where possible URIs were assigned to British Library entities following accepted patterns and best practices (Davidson, 2009).

2.3. The 'Extract, Transform, Load' workflow

To generate Linked Data for the service, relevant BNB MARC21 records are selected from the dataset and passed through a series of character set conversion, data normalisation and matching processes prior to the addition of British Library-minted and external URIs. To place the data in a

[7] https://www.loc.gov/marc/bibliographic/

[8] https://talis.com/ Talis closed its generic semantic web division due to insufficient commercial interest in July 2012

[9] http://www.tso.co.uk/our-expertise/technology/openup-platform

[10] http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf

[11] http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf

[12] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[13] http://www.bl.uk/schemas/bibliographic/blterms

broader context, links to trusted resources selected from both library and general domains are included. The enhanced file is converted to RDF/XML and N-Triples and quality checked. The resulting data dumps are uploaded to the website[14]. N-Triples files together with VoID[15] descriptions are loaded to the Linked Data platform, where users can access the data via a SPARQL endpoint[16] and content negotiation (dereferencing)[17].

3. Data usage analysis methods

3.1. Technology shortcomings

Bringing access analytics to Linked Data requires an understanding of the different modes of Linked Data publication specificities, i.e. dataset dump, SPARQL endpoint and HTTP dereferencing. While monitoring access to a dataset dump is no different from any other file and can be undertaken with traditional Web analytics tools (Fasel and Zumstein, 2009), these applications do not suffice for the two other publication methods. Google Analytics[18] and other popular web analytics platforms[19] (e.g. Open Web Analytics[20], PIWIK[21]) are not designed for linked datasets (e.g. they do not provide access metrics for SPARQL). Getting insights from SPARQL endpoint access requires the parsing of queries issued and extraction of useful information such as the load of a query or the type of resources requested. Similarly, HTTP dereferencing necessitates the support for HTTP 303 content negotiation[17] not handled by traditional Web analytics tools. In the literature, few initiatives propose Linked Data-specific traffic metrics: Möller et al. (2010) propose a list of Linked Data-specific metrics that cover HTTP and SPARQL access to RDF (e.g. ratio between 303 and 200 HTTP requests, number of RDF-aware agents, SPARQL query features, machine vs human classification based on user-agent strings). The well-established USEWOD workshop series[22] is the reference for Linked Data usage mining and provide a dataset of anonymised linked datasets access logs. We reused and extended metrics defined in (Fasel and Zumstein, 2009; Möller et al., 2010).

3.2. System

To assist analysis of the BNB access logs, Fujitsu Ireland developed a hosted analytics platform for Linked Datasets. An online demo[23] shows one month of traffic insights of The British National Bibliography (BNB) data set. The system mines the logs of registered Linked Data publishers and extracts traffic insights. The analytics system is designed for RDF data stores with or without a SPARQL engine, supports Linked Data HTTP dereferencing with HTTP 303 patterns, load-balancing scenarios and filters out search engines and robot activity. The system offers Linked Data-specific features which are currently not supported by classic web analytics tools (e.g. visitor sessions). Clients are not tracked, thus preserving visitors' privacy. To better identify workload peaks of a SPARQL endpoint, SPARQL queries are qualified as heavy or light according to SPARQL syntactic features.

[14] http://www.bl.uk/bibliographic/download.html

[15] https://www.w3.org/TR/void/

[16] http://bnb.data.bl.uk

[17] https://www.w3.org/TR/cooluris/

[18] http://analytics.google.com

[19] https://en.wikipedia.org/wiki/List_of_web_analytics_software

[20] http://www.openwebanalytics.com

[21] http://piwik.org

[22] http://usewod.org/

[23] http://52.49.205.156/analytics/

The system, illustrated in Figure 1, starts by parsing the BNB access logs from March 2014 to April 2015. Access information from robots and search engine crawlers are filtered out to remove noise from usage insights. The system extracts traffic metrics from the logs. It includes traditional metrics such as location of visitors, referrer website as well as Linked Data-specific ones. A list of the metrics extracted is presented in the following section. Traffic metrics are stored in a data warehouse equipped with an SQL-compliant MOLAP[24] unit that answers queries with sub-second latency. The front-end queries the RESTful APIs exposed by the MOLAP Unit, generates a web UI. Figure 2 shows three different screenshots of the Web UI as used by the British Library Linked Data team to get insight on (a) the most popular RDF classes (including visitors' mistakes); (b) the distribution of heavy and light SPARQL queries; and (c) the visitor location over a given period of time. Data can alternatively be accessed directly from the APIs for analysis by other tools or visualisation interfaces. Although the system has been developed in collaboration with the British Library, it is generic and can be used by any Linked Data publisher, providing they have access to their Linked Data server access logs. Distribution plan and licence for the tool can be provided upon request[25].

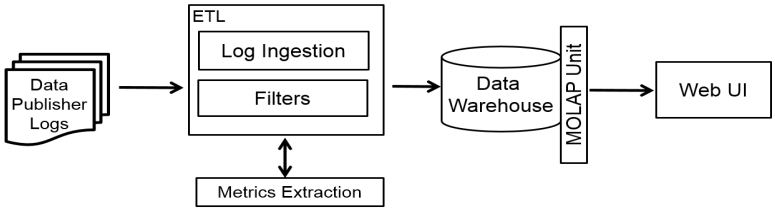
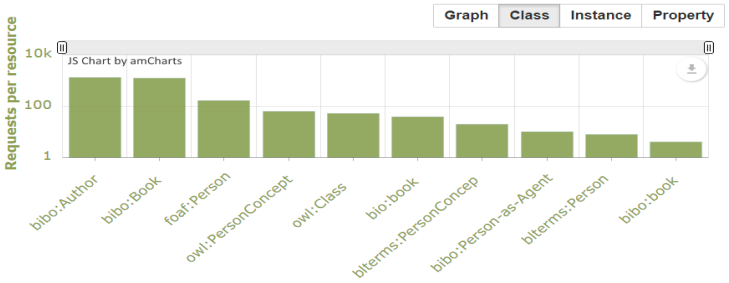
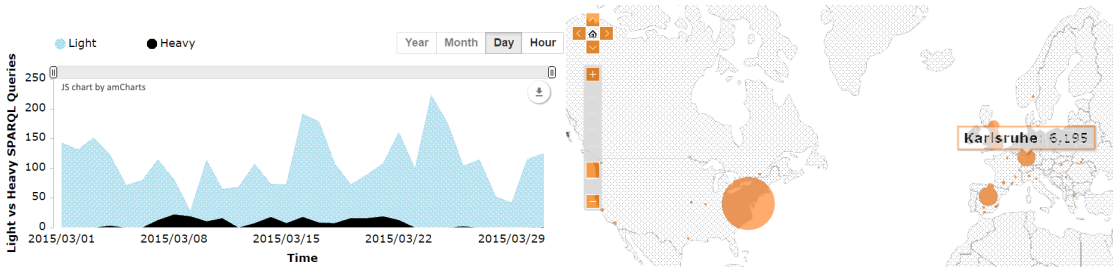


FIG. 1: Architecture of the analytics platform for Linked Data publishers



(a)



(b) (c)

FIG. 2 Screenshots from the web UI (selected timeframe: March 2015): (a) most popular RDF classes (including visitors' mistakes), (b) distribution of heavy and light SPARQL queries, and (c) visitor location.

[24] Multidimensional Online Analytical Processing

[25] http://innovation.ie.fujitsu.com/contact-us/

3.3. Metrics

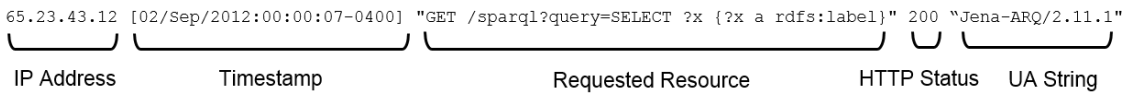


FIG. 3: Linked Dataset access record (Apache Commons Log file Format)

Figure 3 shows an example of access logs processed by the system. A number of key performance indicators for web analytics have been proposed in the academic library domain (Fagan, 2014). The system extracts some of the metrics described by (Fagan, 2014) from access logs, and extends the work of (Möller et al., 2010). Such metrics are grouped in three categories:

Content Metrics. How many times RDF resources have been accessed (cf. Figure 2a). Unlike traditional tools, 303 URIs are correctly interpreted, and the number of times resource URIs appear in SPARQL queries[26] is also counted. Aggregates are provided by family of RDF resource (i.e. instances, classes, properties, graphs).

Protocol Metrics. Information about the data access protocols used by visitors. Includes SPARQL-specific metrics such as the count of malformed queries, SPARQL query type (*SELECT*, *ASK*, *DESCRIBE*, and *CONSTRUCT*) or the detection of light and heavy SPARQL queries (cf. Figure 2b).

Audience Metrics. Besides traditional information about visitors (e.g. location, network provider), these measures include Linked Data-specific metrics such as details of visitor sessions or language tags in queries (cf. Figure 2c).

4. Results

4.1. Overview

While top level statistics for daily and cumulative monthly usage had been regularly recorded (e.g. the BNB dataset dump has been downloaded on average 40 times a month from April 2014 to April 2015), there had been limited detailed examination of log files until the project began. This was due to the volume of data logged, the resources required to extract useful information together with a need to compare data over time to identify meaningful usage patterns. However, the analysis possible via the Fujitsu system offered a range of new insights, which were further assisted by the use of graphical visualisation techniques available on the platform.

4.2. Traffic and usage

As anticipated, the bulk of the search requests (i.e. 43.7M over 13 months) originated from search engine and Linked Data crawlers together with some robot activity. Google variants and other search engines including Bing or Baidu were found to account for 40.7M of these. While significantly smaller as a proportion, the filtered 252K HTTP and SPARQL queries received over the period were found to increase over time from 18K in April 2014 to 24K in April 2015 (see Figure 4) with the number of SPARQL queries making up a significantly increasing proportion (i.e. 67 in April 2014 to 11.1K in April 2015). Over the period of study SPARQL queries were found to be of predominantly light complexity[27] (e.g. 10.8K light vs 364 heavy in April 2015). The average daily duration of sessions[28] was found to be ~1 hour for visitors using software

[26] Access logs do not contain SPARQL result sets. This is therefore a lower bound estimation.

[27] A SPARQL query is defined as heavy (or light) if it requires considerable (or little) computational and memory resources.

[28] A Session is defined as a sequence of requests issued with no significant interruptions by a uniquely identified visitor.

libraries, and 26 minutes for visitors from desktop browsers. Software libraries' sessions also include on average 24 requests with 11 distinct RDF entities queried, while browsers sessions only account for (on average) 2 requests with 2 distinct resources queried. This suggests that although much usage relates to brief investigative, tutorial or test activity (e.g. from desktop browsers), a smaller number of more expert users are undertaking systematic and structured queries of the site over longer periods (e.g. with scripts and SPARQL). This usage breakdown is also evidenced by a 2:1 ratio of new versus returning visitors, indicating the dataset continually receives new users. However, 48% of such sessions consisted of single resource lookups (bounce rate), indicating the possible need to prioritise retention methods, e.g. suggesting "related links" in the HTML view of each RDF resource.

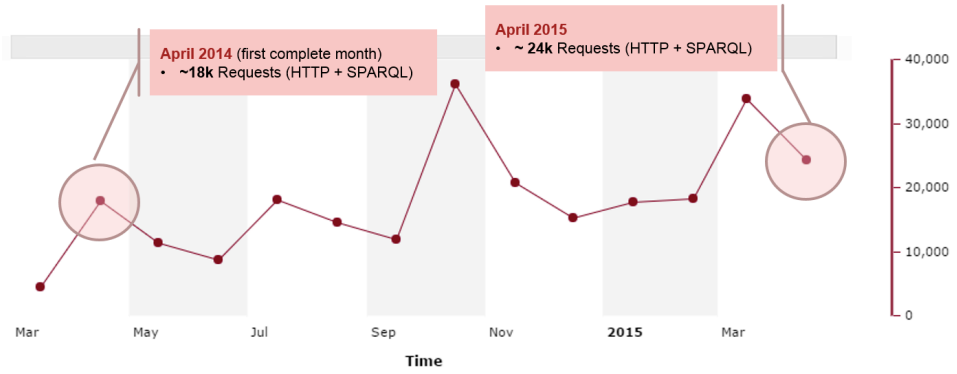


FIG. 4: Evolution of the number of requests from March 2014 to April 2015.

There seemed to be an identifiable correlation between usage peaks and the addition of new metadata elements (e.g. ISNIs[2] in early 2015) or full data refreshes (e.g. September 2014). Some usage patterns could be related to Linked Data tuition sites or events, e.g. a peak in August 2014 following reporting of an international library Linked Data conference in Paris[29] and publication of a Linked Data survey mentioning the BNB[30]. Experimental or educational usage was further underlined by the presence of queries documented in tutorial material published concerning the site, e.g. 1.4K requests for the author C.S. Lewis based on a tutorial query present on the BNB site[31] and 6K requests for 'The Hobbit' relating to a blog tutorial[32]. It was also instructive to quantify previously anecdotal usage by educational institutions. Table 1 presents the ten most queried classes URIs. The fact that several classes queried do not exist in the dataset, e.g. they are misspelled (*bio:birth*), or do not refer to actual terms in existing ontologies (e.g. *owl:PersonConcept*) also reinforces the conclusion that the site is regularly searched by novice users.

TABLE 1: Top 10 classes queried from March 2014 to April 2015

Class URI	Prefixed URI	Present in BNB dataset	Frequency
http://purl.org/dc/terms/BibliographicResource	dcterms:BibliographicResource	Yes	2,115
http://purl.org/ontology/bibo/Author	bibo:Author	**No**	1,429
http://purl.org/ontology/bibo/Book	bibo:Book	Yes	1,307
http://purl.org/vocab/bio/0.1/birth	bio:birth	**No**	591
http://bnb.data.bl.uk/resource/Author	blterms:Author	**No**	476

[29] http://commonplace.net/2014/08/library-linked-data-happening/

[30] http://hangingtogether.org/?p=4137

[31] http://bnb.data.bl.uk/getting-started

[32] https://blog.ldodds.com/2014/10/08/an-introduction-to-the-british-national-bibliography/

http://xmlns.com/foaf/0.1/Person	foaf:Person	Yes	169
http://www.bl.uk/recourse/Author	blterms:Author	**No**	112
http://www.w3.org/2002/07/owl#PersonConcept	owl:PersonConcept	**No**	65
http://www.w3.org/2002/07/owl#Class	owl:Class	Yes	57
http://purl.org/ontology/bio/book	bio:book	**No**	38

The original publication of the BNB Linked Open Dataset generated wide interest amongst global library and open data communities. It was therefore interesting to examine usage from a geographical perspective. The country with the largest source of queries (33.3%) was the United States with the UK next at 21.7% and Germany third at 9.9%. In the library domain, a significant number of requests originated from other state libraries, suggesting a shared exploration of the use of Linked Data at the national library level. The dataset was also found to be used by 350 UK and foreign academic and governmental organisations. The analytics tool also allowed some abnormal activities to be discovered, e.g. an unknown 1-hour spike of 10,000 light SELECT SPARQL queries (October 28th, 2014). The tool identified the requests as originating from a specific city[33], and showed the queries were issued by a Java application. Using this information, staff quickly found thousands of identical queries in access logs and concluded they were probably due to a bug in the client rather than malevolent action.

Queries originating from humans rather than machines accounted for 62% of access with desktop browsers being the most popular method (54%). However, a significant increase (95x) in usage by software libraries was identifiable from 83 requests in April 2014 to 7,895 in March 2015. Overall, there is a clear evolution of the type of visits to access the BNB SPARQL endpoint. This migrated from a dominant profile of manual, human browsing of HTML pages generated from the data (issuing HTTP dereferencing) to a majority of access by machines using software libraries (cf. Figure 5). As of April 2015, 65 distinct SPARQL-based client applications were observed, showing a steady growth from the beginning of the study. This suggests that from an initial experimental base, an ecosystem of more mature clients may be developing around the dataset.

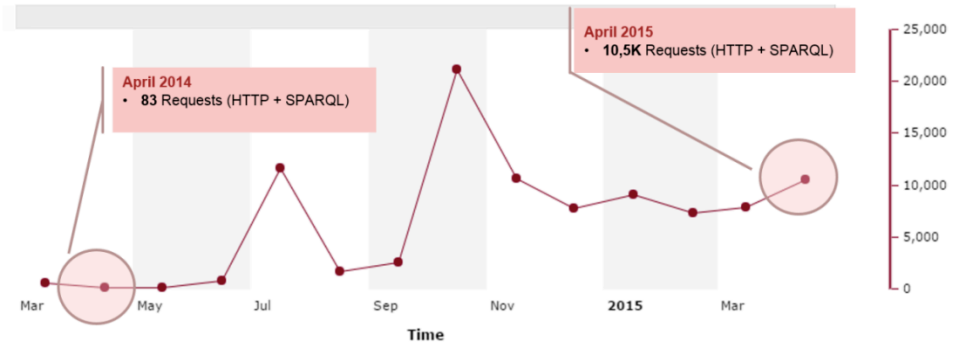


FIG. 5: Evolution of the number of requests issued by Software Libraries.

Client-side HTTP errors account for almost 9% of requests, with 4% of overall requests being '404 Not found' errors (misses). Server-side errors account for 1.5% of total resources (e.g. internal triple store-related errors).

5. Discussion

The role of analytics in assessing Linked Data service value: Creation of comprehensive Linked Data analytics offers Linked Data publishers opportunities to gain new insights into who uses their service and for what purpose and to make development decisions based on concrete

[33] For privacy reasons, we do not provide any specifics about any user.

data. Understanding of BNB LOD usage has significantly improved through the project and its findings will assist future planning. Initially the results have informed development discussions with the current platform provider and will be used to support a case for service continuity based on the identified educational value, the platform's utility for exposing deep library metadata to search engines and indications of a developing Linked Data ecosystem around the service. The results will also be used to guide development and promotion of future Linked Data services and the investigation of a tiered access model to improve performance and prevent misuse. Findings will also assist resource balancing for user documentation activities (e.g. SPARQL/non-SPARQL). Most significantly, the results will inform development of a tender specification for the Library's next Linked Data platform due to go live in 2017.

The wider advantages to publishers offered by Linked Data analytics can be categorised as: organisational, technical, service management, and user support benefits. Some areas, including bounce rate, tracking system performance over time, insights into visitor behaviour and location are common to web analytics; while others have similarities but differing implications. An example of the latter is mobile device usage since this can indicate the impact of a social media campaign or public event via link publication rather than a new requirement for a mobile optimised site.

Organisational benefits: From an organisational perspective, Linked Data analytics offer several important benefits. These include the ability to target scarce resources (staff effort, financial, technical, etc.) more effectively while improving institutional reputation as a trusted creator of LOD services optimised for 'real world' user requirements. The insights gained can also inform an understanding of the relative position of Linked Data platforms in wider institutional systems and resource discovery strategies. Analytics used together with regulated API keys could also assist identification and management of trusted high volume users and potentially lead to new collaboration opportunities.

Technical and service management benefits: Some of the more obvious benefits of Linked Data analytics relate to service management or technical aspects of LOD publication. These include the ability to build a comprehensive understanding of usage categories and geographic spread (e.g. search engine, developer, individual) to support service investment. Such information can support accurate cost control and tendering for service hosting options by ensuring only the appropriate systems capacity is specified. Similarly, such fundamental information enables service providers to determine dataset or feature popularity in order to support accurate decision-making on service extension, enhancement or withdrawal. Specific system benefits from the application of analytics include identification of the range and intensity of normal, legitimate usage together with abuse patterns and abusers to support availability. Normally, publishers have no choice but to manually browse data stored in server access logs. However, an efficient analytics system can extract traffic metrics of Linked Datasets and present results via a web interface to relieve publishers of time-consuming log mining. The ability to interpret access patterns and peak usage can also support service performance optimisation via caching of frequently used data and other tuning techniques. Examination of search engine traffic can also support optimisation of hosted data via targeted monitoring of harvest patterns coupled with structured interrogation of search engine sites to assess the results of changes.

User support benefits: From a support perspective, Linked Data analytics offer staff the ability to identify documentation enhancements (e.g. sample SPARQL queries), relevant tutorial examples and improve evidence-based communications on support issues. An aspect of behaviour common to both web and Linked Data sites is the extremely low percentage of users willing to report problems but to simply switch to alternatives instead. The relatively new nature of LOD services and variations in standards compliance (e.g. SPARQL 1.0/1.1) means user expectations and behaviour can vary significantly beyond those of web equivalents. Regular analytics offer the ability for LOD publishers to spot and fix emerging issues to improve user retention and regular usage while also suggesting development or documentation needs.

All of the above ultimately enable LOD publishers to maintain and improve service continuity and performance for the benefit of users. Concrete analytics data also supports better services with targeted characteristics based on observed usage patterns and developments arising from demonstrable user needs.

6. Conclusions

The British Library believes Linked Open Data to be a logical evolutionary step for the established principle of freedom of access to information, offering trusted and authoritative knowledge organisations an important role in the new information landscape. For such organisations, the vision of a global pool of semantically rich, reusable metadata enabling them to concentrate scarce resources on adding unique value is highly attractive. Similarly, the potential value of LOD sites in offering cost-effective exposure of large datasets to search engines, application developers and new modes of resource discovery has great appeal. However, tough economic conditions and the rapid evolution of LOD solutions necessitate hard evidence-based justification for any new expenditure. The Linked data analytics developed by Fujitsu in this project offers publishers the ability to accurately assess the impact of their data and target scarce resources more effectively. In doing so they can begin to develop and manage new LOD services as efficiently as more traditional web services and continue the realisation of Linked Data's potential for the benefit of the wider community.

Acknowledgements

This work has been supported by the 'TOMOE' project funded by Fujitsu Laboratories Limited.

References

Alemu, Getaneh, Brett Stevens, Penny Ross, and Jane Chandler. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. New Library World, 113(11/12), (pp. 549-570).

Davidson, Paul. (2009). Designing URI Sets for the UK Public Sector. UK Chief Technology Officer Council. Retrieved, May 25, 2016 from https://www.gov.uk/government/publications/designing-uri-sets-for-the-uk-public-sector

Deliot, Corine. (2014). Publishing the British National Bibliography as Linked Open Data. Catalogue & Index (174), (pp. 13-18).

Fagan, Jody Condit. (2014). The suitability of web analytics key performance indicators in the academic library environment. The Journal of Academic Librarianship, 40(1), (pp.25-34).

Fasel, Daniel and Darius Zumstein. (2009). A fuzzy data warehouse approach for web analytics. In Miltiadis D. Lytras et al. (Eds.): WSKS 2009, LNAI 5736, (pp.276-285). Berlin, Heidelberg : Springer.

Hannemann, Jan, and Jürgen Kett. (2010). Linked data for libraries. In Proceedings of the World Library and Information Congress: 76th IFLA General Conference and Assembly, 2010.

Möller, Knud, Michael Hausenblas, Richard Cyganiak, Gunnar Grimnes and Siegfried Handschuh. (2010). Learning from linked open data usage: Patterns & metrics. In WebSci10: Extending the Frontiers of Society On-Line, (pp.1-8)

An Exploratory Study of the Description Field in the Digital Public Library of America

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

Abstract

This paper presents results of an exploratory quantitative analysis regarding the application of a free-text Description metadata element and data values associated with this element. It uses a dataset containing over 11.6 million item-level metadata records from the Digital Public Library of America (DPLA), originating from a number of institutions that serve as DPLA's content or service hubs. This benchmark study provides empirical quantitative data about the Description fields and their data values at the hub level (e.g., minimum, maximum, and average number of description fields per record; number of records without free-text description fields; length of data values; etc.) and provides general analysis and discussion in relation to the findings.

Keywords: metadata aggregations, metadata values, free-text fields, item descriptions.

1. Introduction and Background

Two kinds of metadata coexist in records created according to various metadata standards: controlled-vocabulary metadata which draws values from formally-maintained list of terms, and free-text metadata which relies on natural language. Free-text metadata -- for example, the Description metadata element in the Dublin Core (DC) metadata scheme; various notes (e.g., 5XX fields) in MARC records; Abstract, Note, and Table of Contents elements in the Metadata Object Description Schema (MODS); Scope and Content elements of the Encoded Archival Description (EAD) metadata scheme; etc. -- have been considered an important part of metadata records as a rich source of information on the nature of information object(s) described by each record.

Best practice recommendations have been developed regarding data values for the Description element and its semantic equivalents in metadata records describing information objects -- Cataloging Cultural Objects (CCO) (Baca et al., 2006), Categories for the Description of Works of Art (CDWA) (Baca et al., 2009), OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006) etc. -- as well as in metadata records describing physical collections of manuscripts (National Union Catalog of Manuscript Collections, 2010) and collections of archival materials (OLAC Cataloging Policy Committee, 2002; Encoded Archival Description, 2002, 2015).

Cataloging Cultural Objects (Baca et al., 2006) and Categories for the Description of Works of Art (Baca et al., 2009) suggest recording information about subject, significance, and function in an item-level free-text Description element. OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006) recommends inclusion of provenance and history of the work, as well as the nature of the language of the resource. Dublin Core Usage Guide (Hillmann, 2005) provides guidelines on how to use item-level metadata elements; however, it does not detail what information should be included in Description, besides a broad recommendation, "Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content" (section 4.3).

Several documents discuss specific guidelines in relation to collection-level metadata rather than item-level records. National Union Catalog of Manuscript Collections (2010) suggests that collection-level metadata creators for manuscript collections provide in the Description element: information about types of materials included in the collection; topics with which the materials in the collection deal; geographical areas with which the materials in the collection deal; associated dates, events, and historical periods dealt with by the materials in the collection; names, dates, and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection; and specific phases of career/activity of the major person or corporate body responsible. Summary Notes for Catalog Records (OLAC, 2002) recommends inclusion of information about specific types and forms of materials present; significant people, topics, places, and events covered; span of dates covered by the collection; history of the work; unique characteristics of the collection; reason and function of the collection; audience; and user interaction. The previous version of Encoded Archival Description (EAD, 2002) recommended inclusion of such characteristics as form and arrangement of materials; significant subjects represented; places represented; events represented; significant organizations and individuals represented; and collection strengths. The current version of Encoded Archival Description (EAD3, 2015) adds a recommendation to provide information functions and activities that generated the materials being described, and gaps in the materials to help the user evaluate the potential relevance of the materials being described.

The guidelines on constructing data values for free-text metadata elements, such as Description, are intended to facilitate users' access to information objects and collections through these rich metadata fields, but the suggestions are not necessarily followed in creation of metadata records. Empirical research focusing on analysis of data values in free-text Description metadata allows researchers not only to determine the level of adherence to guidelines but also, importantly, to categorize information typically found in these data values. For example, two studies of collection-level metadata in large-scale repositories in the United States and Europe (Zavalina et al, 2008; Zavalina, 2012) resulted in the list of 19 properties of a digital collection that are represented in Description fields in collection-level metadata records: topical coverage; geographic coverage; temporal coverage; collection title; size; collection development information; provenance; importance of collection; uniqueness; comprehensiveness; intended audience; navigation and functionality; participating, hosting or contributing institutions; copyright information; frequency of additions to collection; funding sources; types/genres of items; creators of items; and language of items. Several of these categories of information (e.g., creators of items in collection, etc.) did not appear in the existing guidelines for the free-text Description field but were nevertheless included by metadata creators who considered these important for more efficient information access and discovery.

For item-level metadata, several studies looked at frequency of application of Dublin Core metadata elements, including the free-text Description element, in metadata aggregations. For example, in Ward's (2003) study of over 900,000 Dublin Core metadata records harvested from 82 OAIster data providers, it was observed that Description element was included in slightly over a half (50.9 %) of all records and that 72% of data providers included this element in their records. Jackson and colleagues (2008), in their study of metadata harvested into IMLS DCC aggregation , did not report the observed percentage of metadata records that include Description field, but reported its systematic inclusion in the records from 31 (89%) out of 35 harvested digital collections. The findings of these studies show the level of application of Description elements to range substantially.

Other studies measured application of metadata elements, including Description element in DC-based metadata and/or its counterparts from other metadata schemes in individual digital collections. For example, Kurtz's (2010) analysis of metadata applications in three digital repositories hosted by university libraries and using Dublin Core demonstrated fluctuations in the level of Description element usage, from 40% to 75% of metadata records. A study of metadata application in digital video collections (Weagley, Gelches, & Park, 2010) revealed a much higher

level of application of Description metadata element (99% of metadata records, the highest of all elements and at the same level with the Title element) than other studies that measured application of Dublin Core metadata. This might be due to the specific nature of these digital collections. Similar observations were made for three digital image collections in a study (Park, 2006) which found the Description element to be included in all 100% of Dublin Core metadata records across the collections.

1.1. Digital Public Library of America

The Digital Public Library of America (DPLA) is a prominent aggregation of metadata, currently comprising over 13 million metadata records from libraries, archives and museums in the United States to provide free public access. DPLA functions on a distributed network model and consists of a group of national partners or "hubs" providing both content and services (Ma, 2014). Content hubs constitute large libraries, museums, archives and other digital repositories which maintain a one-to-one relationship with DPLA. Service hubs are state, regional, or other collaborations which host, aggregate, or otherwise bring together digital objects from cultural heritage institutions and provide metadata to the DPLA through a single data feed such as OAI-PMH.

The internal data model of DPLA is based on the Resource Description Framework (RDF) and the central descriptive metadata standard employed is the Dublin Core (Mitchell, 2013). In DPLA, some of the metadata gathered from providers is stored along with metadata generated or extracted during the aggregation process. The metadata aggregated and normalized by DPLA is in the public domain and has no copyright restriction; DPLA metadata can be harvested via the OAI-ORE standard for sharing or data analysis. JSON-LD (JavaScript Object Notation-based serialization for Linked Data), an RDF-inspired serialization, is disseminated via API output.

In the most recent version of DPLA metadata documentation, there is an inconsistency regarding the status of the Description property: in the Introduction to version 4 of the DPLA metadata model (Digital Public Library of America, 2015a, p.9), the Description property of the sourceResource class is named a "recommended" metadata element -- i.e., an element that should be included in a metadata record if the information is available -- but in the complete DPLA Metadata Application Profile document (Digital Public Library of America, 2015b, p.20), this property is not included in the listing of required or recommended properties. In DPLA's metadata application profile, which is based on an RDF serialization of the Dublin Core descriptive metadata standard, the DPLA Description element maps to dcterms:description (Digital Public Library of America, 2015a). Native metadata -- metadata used internally by institutions that serve as DPLA hubs -- is often more detailed and relies on richer metadata schemes than Dublin Core, such as MODS or MARCXML. Multiple metadata elements from these metadata schemes (e.g., MODS abstract, tableOfContents, and note; various 5XX MARC fields; etc.) map to a single metadata element (Description) in Dublin Core. Therefore, as a result of normalizing and aggregating native metadata into DPLA, it is likely that metadata records contain multiple Description fields with varying kinds of data values.

The review of the literature demonstrates the lack of recent empirical, quantitative studies of free-text description metadata. The study reported in this paper is one of the first attempts to systematically evaluate this kind of metadata, and the first one to use a very large aggregator such as Digital Public Library of America as its target.

2. Methods

The research questions that guided this exploratory study fall into two areas: (1) What is the overall usage of the Description field by hubs in the DPLA dataset? And (2) How can high-level attributes such as length of data values provide insight into metadata practices regarding the free-text Description metadata field among DPLA hubs?

To address these research questions, we applied the quantitative content analysis research method. Unlike many previous studies of metadata in large-scale digital libraries that analyzed a generalizable sample of metadata records, the authors of this study took a "big data" approach that analyzes the whole dataset and therefore avoids sampling errors. The authors used DPLA's Bulk Download to harvest the metadata dataset (http://dp.la/info/developers/download/). This dataset was parsed into individual records that contain both the original metadata submitted by various DPLA hubs and a normalized version based on the DPLA Metadata Application Profile (http://dp.la/info/developers/map/).

For this analysis, each record was parsed from the DPLA dataset and processed to extract the Description field information, along with the DPLA identifier for the record and the originating provider/hub. The resulting dataset comprises 11,654,800 records. Because the Description field is not required and is repeatable, some records contain no Description values while other records contain multiple instances of the Description field. The original 11,654,800 records in the DPLA dataset contained a total of 17,884,946 description values. Each record was further processed to generate metrics about individual Description field instances. Examples of these metrics include: length of description (number of characters); number of words; average word length; and proportion of description that consists of letters, punctuation, or integers. In total there were 20 descriptive metrics generated for each of the description values in the dataset.

3. Findings

All of the Description field values were loaded into the Apache Solr Full-Text indexer where various components of that system including the facet and the statistics components were used to explore the dataset.

For each analysis, the findings were broken down by hub. A relatively small number (11,422) of records did not include hub source information; for the purposes of maintaining completeness of the dataset, these are categorized as records originating from "undefined provider."

3.1. Usage

The first general analysis included a count of instances of Description values per record (Table 1). Since this field is repeatable and serves as point to which many free-text fields map from the hubs, some records have more than one instance of the Description field.

As shown in Table 1, there is a wide range of usage in the Description field across hubs. In some cases, a large majority of records have no Description field values. These include collections from the National Archives and Records Administration (NARA, 98.83%), Kentucky Digital Library (98.66%), and items with undefined provider (99.89%). On the other end of the spectrum, The Portal to Texas History includes Description fields in 99.98% of its metadata records and several others -- the United States Government Publishing Office (GPO), J. Paul Getty Trust, David Rumsey, and University of Illinois at Urbana-Champaign -- also have at least one Description field value in more than 99% of their records.

The number of Description instances per record also represents a drastic range (see Fig. 1). Eight hubs -- Biodiversity Heritage Library, Empire State Digital Network, Kentucky Digital Library, Minnesota Digital Library, NARA, Tennessee Digital Library, University of Virginia Library, and University of Washington -- have no more than one Description value in any record (see Appendix A for additional statistics). However, some item records contain an extremely large number of values. The Smithsonian Institution has at least one record containing 179 separate Description entries; the Digital Library of Georgia and Indiana Memory each have at least one record with 98 separate entries. While these numbers seem to be outliers on the whole, five other hubs have records containing at least 25 separate Description values: HathiTrust (77), GPO (65), Internet Archive (35), J. Paul Getty Trust (25), and University of Illinois at Urbana-Champaign (25).

Additionally, our analysis considered the total number of Description field instances in metadata records per hub, as well as the percentage of those Description field data values that are unique (Table 1). The three hubs that have less than 1% uniqueness are the same hubs that have few Description field instances in their records: Kentucky Digital Library, NARA, and undefined provider. This suggests that the few records that *do* contain Description field values from these hubs have significant content overlap.

TABLE 1: Distribution of Description field instances in metadata records by hub.

Hub	Records	Records with 0 Description Instances		Records with 1+ Description Instances		Total Instances	Unique Description Values	
artstor	107,665	40,851	37.94%	66,814	62.06%	128,922	34,490	26.75%
bhl	123,472	64,928	52.59%	58,544	47.41%	123,472	46,235	37.45%
cdl	312,573	80,450	25.74%	232,123	74.26%	563,967	300,983	53.37%
david_rumsey	65,244	168	0.26%	65,076	99.74%	166,314	32,093	19.30%
digital-commonwealth	222,102	8,932	4.02%	213,170	95.98%	455,369	110,200	24.20%
digitalnc	281,087	70,583	25.11%	210,504	74.89%	241,224	162,178	67.23%
esdn	197,396	48,660	24.65%	148,736	75.35%	197,396	91,001	46.10%
xgeorgia	373,083	9,344	2.50%	363,739	97.50%	821,067	271,437	33.06%
getty	95,908	229	0.24%	95,679	99.76%	264,268	32,419	12.27%
gpo	158,228	207	0.13%	158,021	99.87%	690,883	208,307	30.15%
harvard	14,112	3,106	22.01%	11,006	77.99%	23,645	14,487	61.27%
hathitrust	2,474,530	1,068,159	43.17%	1,406,371	56.83%	4,077,994	1,449,785	35.55%
indiana	62,695	18,819	30.02%	43,876	69.98%	74,009	35,907	48.52%
internet_archive	212,902	40,877	19.20%	172,025	80.80%	521,102	128,870	24.73%
kdl	144,202	142,268	98.66%	1,934	1.34%	144,202	693	0.48%
mdl	483,086	44,989	9.31%	438,097	90.69%	483,086	195,321	40.43%
missouri-hub	144,424	17,808	12.33%	126,616	87.67%	169,332	89,907	53.10%
mwdl	932,808	57,899	6.21%	874,909	93.79%	1,195,954	741,141	61.97%
nara	700,948	692,759	98.83%	8,189	1.17%	700,948	4,667	0.67%
nypl	1,170,436	775,361	66.25%	395,075	33.75%	1,170,438	61,423	5.25%
scdl	159,092	33,036	20.77%	126,056	79.23%	159,598	53,974	33.82%
smithsonian	1,250,705	68,871	5.51%	1,181,834	94.49%	2,805,327	343,372	12.24%
the_portal_to_texas_history	649,276	125	0.02%	649,151	99.98%	1,271,500	234,696	18.46%
tn	151,334	2,463	1.63%	148,871	98.37%	151,334	129,605	85.64%
uiuc	18,231	127	0.70%	18,104	99.30%	63,403	25,123	39.62%
undefined_provider	11,422	11,410	99.89%	12	0.11%	11,436	16	0.14%
usc	1,065,641	852,076	79.96%	213,565	20.04%	1,076,016	182,084	16.92%
virginia	30,174	21,081	69.86%	9,093	30.14%	30,174	1,118	3.71%
washington	42,024	8,838	21.03%	33,186	78.97%	42,024	20,710	49.28%

Among larger collections, however, the amount of duplication in Description values does not follow similar patterns. The four hubs containing more than 1 million items -- HathiTrust, New York Public Library, the Smithsonian Institution, and University of Southern California Libraries -- have uniqueness values ranging from a mere 5.25% to nearly 36%. In addition to the four largest contributors, two other hubs have more than 1 million descriptions, though fewer items: The Portal to Texas History (1,271,500 descriptions with only 18.5% uniqueness) and Mountain West Digital Library (1,195,945 descriptions with roughly 62% uniqueness). Tennessee Digital Library has the highest level of uniqueness (86%) with only 151,334 items. Overall, there do not appear to be any generalizable correlations among collection size, number of descriptions, and uniqueness.

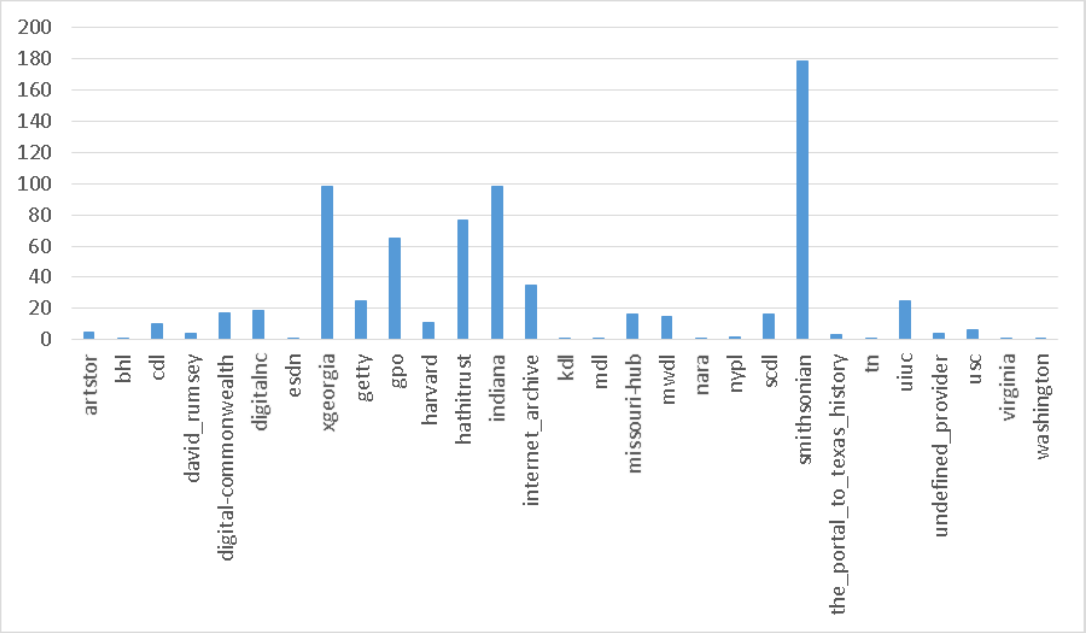


FIG. 1. Largest number of Description instances in any record, by DPLA hub.

3.2. Description Length

After looking at usage of the Description field by hubs, we wanted to gain a better sense of the field values and compare them across the dataset.

Our analysis shows that the length of Description field values in all DPLA metadata records averaged 83.3 words. The range of Description lengths was very broad, with a standard deviation of 373.71 and a maximum length of 130,592 words (approximately 45 pages of text).

Table 2 shows the distribution of Description value lengths by hub. Our analysis identified five hubs with the highest average lengths, ranging from 201 to 447 words: David Rumsey, J. Paul Getty Trust, Minnesota Digital Library, Missouri Hub, and Tennessee Digital Library. On the other side of the spectrum, three out of five hubs with the shortest average length of data values (i.e., under 10 words) are the same three hubs with the lowest number of records containing Description fields and the lowest level of uniqueness: Kentucky Digital Library (2.71 words), NARA (2.03 words) and undefined provider (0.21 words). The other two hubs with the shortest lengths of data values are Biodiversity Heritage Library (6.29 words) and University of Virginia Library (9.98 words).

It is also notable that the spread of lengths is vast for some hubs, e.g., Missouri Hub with an average of 210 characters, but a standard deviation of 2325. Mountain West Digital Library and David Rumsey both have extremely large standard deviations also, with 905.5 (average 154.6 characters) and 861.92 (average 447.36 characters) respectively. The smallest standard deviation (aside from "undefined provider") is Biodiversity Heritage Library (8.48), though the average length is only 6.28 characters.

Figure 2 shows lengths of Description values on a log-log scale. A noticeable spike at 10 characters sets off the group of extremely short descriptions. Although 4.1 million records have no Description values (i.e., a length of 0), they do not display on the log scale; the set from 1-10 characters is more than 2 million descriptions (roughly 2%). On the far left axis, nearly 800,000 values are only a single character long. From that point, the graph shows a clear inverse relationship between the number of characters and the number of records in which they appear (i.e., records with larger numbers of characters tend to occur less frequently). However, there are several obvious spikes, particularly around 800-1,000 characters and 1,500-1,800 characters. These longer values likely represent full sentences and paragraphs rather than the single or few words in the shorter values.

TABLE 2: Description field length statistics by hub.

Hub	Minimum Length	Maximum Length	Instances	Sum of Lengths	Mean/ Average	Standard Deviation
artstor	0	6,868	128,922	9,413,898	73.02	178.31
bhl	0	100	123,472	775,600	6.28	8.48
cdl	0	6,714	563,967	65,221,428	115.65	211.47
david_rumsey	0	5,269	166,314	74,401,401	447.36	861.92
digital-commonwealth	0	23,455	455,369	40,724,507	89.43	214.09
digitalnc	0	9,785	241,224	45,759,118	189.66	262.89
esdn	0	9,136	197,396	23,620,299	119.66	170.67
xgeorgia	0	12,546	821,067	135,691,768	155.05	210.85
getty	0	2,699	264,268	80,243,547	303.64	273.36
gpo	0	1,969	690,883	33,007,265	47.81	58.20
harvard	0	2,277	23,645	2,424,583	102.54	194.02
hathitrust	0	7,276	4,077,994	174,039,559	42.66	88.03
indiana	0	4,477	74,009	6,893,350	93.93	189.30
internet_archive	0	7,685	521,102	41,713,913	79.68	174.94
kdl	0	974	144,202	390,829	2.71	24.95
mdl	0	40,598	483,086	105,858,580	219.13	345.47
missouri-hub	0	130,592	169,332	35,593,253	210.14	2325.08
mwdl	0	126,427	1,195,954	174,126,243	145.60	905.51
nara	0	2,000	700,948	1,425,165	2.03	28.13
nypl	0	2,633	1,170,438	48,750,103	41.65	161.88
scdl	0	3,362	159,598	18,422,935	115.37	164.74
smithsonian	0	6,076	2,805,327	139,062,761	49.52	137.37
the_portal_to_texas_history	0	5,066	1,271,500	132,235,329	104.00	95.95
tn	0	46,312	151,334	30,513,013	201.63	248.79
uiuc	0	4,942	63,403	3,782,743	59.65	172.44
undefined_provider	0	469	11,436	2,373	0.21	6.09
usc	0	29,861	1,076,016	60,538,490	56.26	193.20
virginia	0	268	30,174	301,042	9.98	17.91
washington	0	1,000	42,024	5,258,527	125.13	177.40

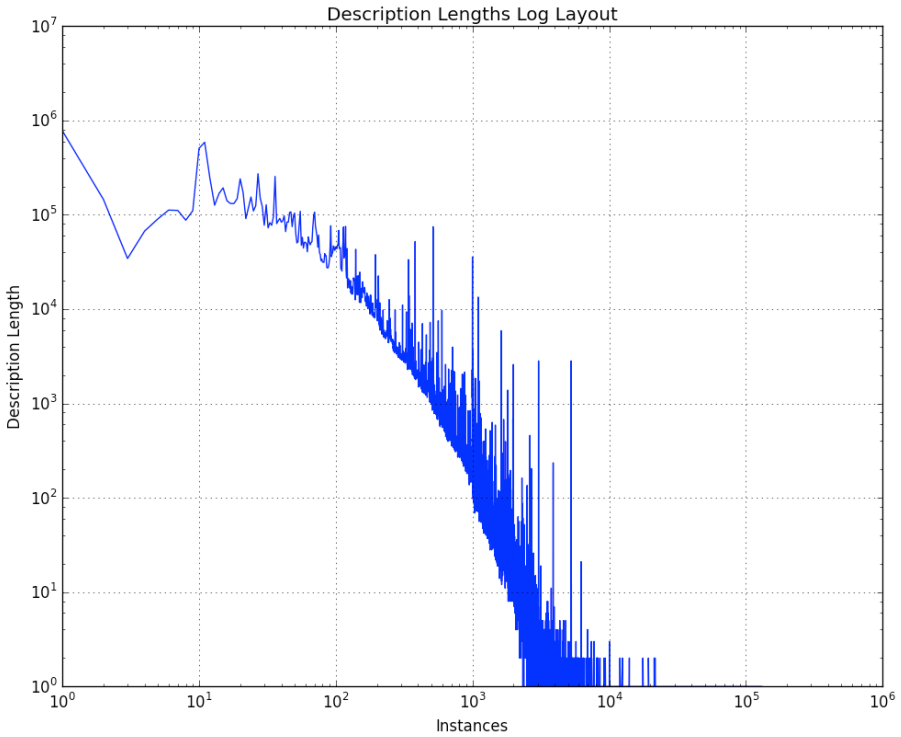


FIG. 2. Lengths of all Description values in the DPLA dataset.

4. Discussion

Although the amount of information gleaned from this form of analysis can be somewhat limiting, there are some definite points that we can make based on our observations, and in the course of considering our research questions with the data in mind.

First, we wanted to look at overall usage of the Description field among DPLA hubs. The number of records without Description values suggests that different hubs (and perhaps different partner institutions, in the case of service hubs) may not all consider the Description field to be equally important, or may not enforce the usage of a Description field. However, considering DPLA's practices, which map many descriptive or "note" type fields that do not map elsewhere into the Description field, it appears that many hubs do not allow or actively record descriptive information of any kind at an item level. This is somewhat at odds with the literature, which suggests the use of free-text fields as a way of most adequately imparting important information to users about the items. Comparing perceived importance with actual usage could be an interesting source of potential future research.

Our other initial question sought to discover what we might learn about usage of the Description field (among records containing Description values) by looking at their various attributes. The most obvious pattern is that there is essentially no pattern -- the number of element instances and the length of field data values vary wildly across DPLA contributing hubs and, in some cases, within hub collections. However, we do wish to offer some additional explanations about why this emerged in the data and what it could mean going forward.

4.1. Description Lengths

Aside from outliers, it is hard to draw definitive conclusions about the range of lengths. For example, longer Description field lengths could indicate more rigorous description standards in these hubs (e.g., specific guidelines on the level of detail that should be included in Description field values). The lengths could also be dependent on specific description practices for the types of information objects that are prevalent in collections of these hubs (e.g., image-based materials may have longer descriptions than collections of primarily printed text with OCR files). Exploring these aspects could be a useful opportunity for future research.

In addition to the large data analysis, we surveyed a random sample of roughly 200 Description field data values per hub for a total of 2800 data values. Although this sample is not very large and we did not have enough time to draw definitive conclusions within the scope of this paper, the Description values do provide some insight into the statistics and allow us to make more educated evaluations of the previous analysis (see Table 3).

TABLE 3: Selected Description field values from DPLA records.

Description Value	Information Type
1 glass negative: b&w; 8 x 10 in.; sulfiding.	Physical object description
This material has been provided by The Royal College of Surgeons of England. The original may be consulted at The Royal College of Surgeons of England	Rights or usage statement
This image shows a section of Thorn Cemetery including gravestones.	Object content description
Microform.	Object type or format
Title supplied by cataloger.	Note or metadata source
This series contains transcripts of proceedings, depositions, and oral examinations prepared exclusively for or in the District Court. The depositions and oral examinations were taken out of court and are primarily interviews with School Board representatives and employees concerning the development, implementation, and review of desegregation plans.	Collection-level content description
P950.	Identifier or call number

The sample includes data values containing a variety of information, such as rights and use statements, physical descriptions, and collection-level descriptions. These kinds of descriptions

may account, in part, for the relatively high number of duplicate values observed in the data. For example, if the same use statement or collection information is propagated across a large group of item records, it would reduce the number of unique data values in a collection of metadata records. Similarly, records may have the same Description field values describing physical attributes that are identical for many items (e.g., 13 p. or 4 x 5 in.), or describing the content attributes of a large number of serial items.

This also provides one explanation for the large number of extremely short Description values. Some hubs use Description fields to hold data values that contain only an item count, page count, or a short term from a controlled vocabulary. In addition, some Description values identify names of places, people, or events without contextual information, which likely accounts for some shorter data values.

4.2. Mapping

Since DPLA is an aggregation, much of the information available in these records is in a shortened format depending on how it is harvested or the level of normalization to fit the DPLA profile. As a relatively generic, free-text field (which also has no strict guidance or recommendations), Description serves as a mapping point for many different native metadata fields. This also makes it difficult to determine if the variety of information types observed in the dataset analyzed in this study is due to differing perceptions of "Description" among hubs and contributors, if there is simply no better place to map the information in DPLA, if the contributed records are too inconsistent to map more accurately, or some combination of all of these factors. However, it does seem that some information found in DPLA metadata records' Description fields could/should be mapped to a more appropriate field (e.g., rights statements).

This is another area that could benefit from much deeper research in terms of how different institutions define or perceive item-level (and collection-level) metadata, both in native systems and as part of an aggregation. Additional research may also consider classifying values currently mapped to Description and the possibility of automatically identifying some information to map values more accurately or to mark them for review for quality control.

4.3. Context and Quality

While not conclusive, several of the statistics identified within this research can help identify metadata records within the DPLA dataset that are in need of remediation. Specifically, records that have Description field values of more than 20,000 characters should be reviewed as to their appropriateness to local descriptive metadata input rules. In many cases, the values at the high end of the length spectrum likely contain the full text of the materials described by the records, and suggests possible problems with the quality of metadata records.

At the same time, records with extremely short values suggest the need for additional review in order for users to understand the information in its aggregated form. Institutions could consider a change in the way that the data values are entered, if one of the primary goals for those institutions is to make information shareable/aggregatable (thought it may not be). Aside from local changes, perhaps there is some potential for preserving or representing more of the contextual information that has been lost within the aggregation.

Even in a native system, extremely short descriptions that are part of a free-text field may suggest a lack of relevant information about the item. For example, a three-word description, such as "A view east" could be accurate in relating to an item without providing sufficient context to help users understand an item's relevance; this statement could refer to a photograph (of nearly anything), a poem title, a map, etc. Similarly, although identifying a name or location is generally considered important, without any context, a proper name remains extremely vague -- e.g., is the name of a person describing an individual pictured in a photo or artwork, a donor, one person in a group photo, or the subject of an obituary or text? From this perspective,

contextual information within a Description or free-text field could be considered highly important to the quality of the field value and the metadata record's usefulness.

5. Conclusions

The empirical data collected and analyzed in this study allows us to make a conclusion that simple statistical analyses can provide a better understanding field usage within a large metadata set. In this case, by investigating the Description fields from the Digital Public Library of America, we were able to consider a wide range of conceptual and technical models for metadata creation by a large number of institutions across the country. This diversity allows for a better understanding of practices than similar analysis within a single institution. However, our findings also show that the Description field and the nature of aggregated free-text fields are areas that would greatly benefit from additional research that was outside our scope and time constraints.

5.1. Further Research

This research was not able to take advantage of the majority of the Description attributes indexed in the methods described above. Performing similar analysis on these additional attributes would result in a better understanding of how the Description field is being used at a wide range of institutions, beyond the usage and length metrics.

Some areas of specific interest for further research include the use of language by each of the providers. This was calculated by identifying, for each of the Description values, the percentage of words that come from various lists of frequently-used English words (e.g., comparing data values to the 1,000 and 5,000 most frequently used English words, and against a standard English dictionary). Additionally, further investigation in this area could provide insight into the reading levels and intended audiences of the metadata being created at each of the provider/hubs. Along these same lines, research into how descriptive information helps users find items and the perception of usefulness by user communities could help to refine guidelines around Description field usage and importance.

On a broader level, the analysis in this report represents a "distant reading" of metadata values in a large dataset. In order to further understand the use of the Description field in the DPLA metadata aggregation, a "close reading" of the Description field values would be beneficial to practitioners and technologists working with metadata aggregations.

References

Baca, M. and P. Harpring (Eds.). (2009) Categories for the Description of Works of Art (CDWA), Getty Research Institute, Santa Monica.

Baca, M., et al. (2006) Cataloging Cultural Objects: A Guide to Describing Cultural Works and their Images, American Library Association, Chicago.

Digital Public Library of America (2015a, March 5). An introduction to the DPLA metadata model. Retrieved from http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf

Digital Public Library of America (2015b, March 5). Metadata application profile: Version 4.0. Retrieved from http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf

Encoded Archival Description. (2002). Retrieved from http://www.loc.gov/ead/ .

Encoded Archival Description: EAD3. (2015). Retrieved from http://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf.

Jackson, A.S., M. Han, K. Groetsch., M. Mustafoff and T. W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8*(1), 5-21.

Hillmann, D. (2005). Using Dublin Core. Retrieved from http://dublincore.org/documents/usageguide/

Kurtz, M. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology & Libraries, 29*(1), 40-46. Retrieved from http://ejournals.bc.edu/ojs/index.php/ital/article/view/3157/2771

Ma, Hong. (2014). Techservices on the Web: DPLA: Digital Public Library of America. *Technical Services Quarterly, 31*(1), 83-84. doi: 10.1080/07317131.2014.845013

Mitchell, Erik T. (2013). Three case studies in linked open data. *Library Technology Reports*, *49*(5), 26-43.

National Union Catalog of Manuscript Collections. (2011). Online Data Sheet for Participating Institutions. Retrieved from http://www.loc.gov/coll/nucmc/lcforms.html.

OLAC Cataloging Policy Committee, Summary/Abstracts Task Force. (2002) Summary Notes for Catalog Records. Retrieved from http://www.olacinc.org/drupal/?q=node/21.

OSU Knowledge Bank Metadata Application Profile for Digital Video. (2011). Retrieved from https://library.osu.edu/documents/knowledge-bank/KnowledgeBankMetadataApplicationProfile2011.pdf

Park, J. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge Organization*, *33* (1), 20-34.

Ward, J. (2003). A quantitative analysis if unqualified Dublin Core metadata element set usage within data providers registered with the Open Archives Initiative. Proceedings of the 2003 Joint Conference on Digital Libraries, pp. 315-317.

Weagley, J., E. Gelches, & J. Park. (2010). Interoperability and metadata quality in digital video repositories: a study of Dublin Core. *Journal of Library Metadata*, *10*(1), 37-57. DOI: 10.1080/19386380903546984.

Zavalina, O.L. (2012). Exploring the richness of collection-level subject metadata in three large-scale digital libraries. *International Journal of Metadata, Semantics, and Ontologies*, *7*(3), 209-221.

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

Appendix A

TABLE 4: Distribution and statistics for Description field instances in metadata records by hub.

Hub	Records	Minimum	Median	Maximum	Average	Standard Deviation
artstor	107,665	0	1	5	.82	.84
bhl	123,472	0	0	1	.47	.50
cdl	312,573	0	1	10	1.55	1.46
david_rumsey	65,244	0	3	4	2.55	.80
digital-commonwealth	222,102	0	2	17	2.01	1.15
digitalnc	281,087	0	1	19	.86	.67
esdn	197,396	0	1	1	.75	.43
xgeorgia	373,083	0	2	98	2.32	1.56
getty	95,908	0	2	25	2.75	2.59
gpo	158,228	0	4	65	4.37	2.53
harvard	14,112	0	1	11	1.46	1.24
hathitrust	2,474,530	0	1	77	1.22	1.57
indiana	62,695	0	1	98	.91	1.21
internet_archive	212,902	0	2	35	2.27	2.29
kdl	144,202	0	0	1	.01	.12
mdl	483,086	0	1	1	.91	.29
missouri-hub	144,424	0	1	16	1.05	.70
mwdl	932,808	0	1	15	1.22	.86
nara	700,948	0	0	1	.01	.11
nypl	1,170,436	0	0	2	.34	.47
scdl	159,092	0	1	16	.80	.41
smithsonian	1,250,705	0	2	179	2.19	1.94
the_portal_to_texas_history	649,276	0	2	3	1.96	.20
tn	151,334	0	1	1	.98	.13
uiuc	18,231	0	3	25	3.47	2.13
undefined_provider	11,422	0	0	4	.00	.08
usc	1,065,641	0	0	6	.21	.43
virginia	30,174	0	0	1	.30	.46
washington	42,024	0	1	1	.79	.41

Permanence and Temporal Interoperability of Metadata in the Linked Open Data Environment

Shigeo Sugimoto
University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp

Chunqiu Li
University of Tsukuba, Japan
licq.chunqiu@gmail.com

Mitsuharu Nagamori
University of Tsukuba, Japan
nagamori@slis.tsukuba.ac.jp

Jane Greenberg
Drexel University, USA
jg3243@drexel.edu

Abstract

This paper examines metadata longevity issues in the Linked Open Data (LOD) environment, where metadata, as a digital object, is transferred and shared on the open Web. Longevity is key for achieving metadata permanence, which allows metadata to remain interpretable by machines and humans over time. The discussion presented in this paper seeks to clarify risks in permanence of metadata, by focusing on metadata longevity challenges specific to metadata and metadata schemas in the LOD environment.

This examination addresses metadata longevity from several different viewpoints in order to clarify the requirements of metadata permanence in the LOD environment, and distinguish these needs from conventional document-like object environment or database-centric environment. A central theme in this work is that longevity of metadata is, in essence, the temporal interoperability of metadata. This paper uses the Metadata Application Profile methodology supported by the Dublin Core Metadata Initiative (DCMI) and DCMI's layered model of metadata interoperability to understand the nature of metadata in the LOD environment. Next, the paper discusses metadata longevity based on a set of facets of metadata entities such as metadata schemas; and the last part briefly discusses issues to use provenance description of metadata schemas and metadata schema registries from the viewpoint of long-term maintenance of metadata schemas.

Keywords: metadata longevity; digital preservation; provenance description; metadata schema registry; metadata schema maintenance

1. Introduction

The importance of metadata for digital object preservation is well recognized with metadata standards, such as PREMIS and METS, both supporting property sets for document object preservation. The longevity of metadata objects in the Web environment is, however, still largely unexplored. This is because metadata schemas are treated as instances conventional, operational database systems, or as document like objects, with longevity challenges addressed through common document or database preservation methods. In comparison, metadata in the open Web environment has different features and functionalities. Web resources may include metadata embedded in the headings and/or bodies; and there are many LOD conformant datasets available on the Web that may be used to link Web resources and other objects. A significant feature of metadata on the Web is that metadata instances may be transferred from site to site and saved for the future use. For example, metadata embedded in a HTML header and Cataloging In Publication (CIP) included in a digital book may be extracted and transferred as a digital object. Another example is RDF encoded metadata instances that can be downloaded from LOD datasets for different applications. A metadata schema, which defines structural, syntactic and semantic features of metadata, can also be transferred on the Web as well as the metadata instances because they are metadata about metadata. In fact, metadata transferrable as a digital object on the Web is First Class Object, as examined in this paper.

DCMI Application Profiles (DCAP) provide an intellectual and structural framework for mixing-and-matching metadata vocabularies to define a metadata schema for an application. LOD recommends the use of standards such as OWL and RDF to define metadata schemas and vocabularies in order to make metadata interoperable. These conventions and best practices present a significant challenge when considering metadata longevity. On one hand, metadata application profiles use well-standardized, mature and fairly stable metadata vocabularies. On the other hand, the schema may rely on many components defined outside of the immediate schema. Changes to any of these external components can have a significant impact on the application schema's functionality.

Leading researchers have been hosting two metadata schema registries – DCMI Metadata Registry and MetaBridge[1]. These registries are developed based on RDF and LOD technologies. The DCMI Registry is dedicated to providing access to DCMI terms (Nagamori, Baker, Sakaguchi, Sugimoto, & Tabata, 2001). MetaBridge provides functions to store and provide application profiles in addition to metadata vocabularies and terms (Nagamori, Kanzaki, Torigoshi, & Sugimoto, 2011). These registries function differently than data and resource repositories. That is, they are generally not developed for long-term maintenance of metadata schemas but they have a large potential to serve as a long-term maintenance host, and could in fact be developed to address similar goals.

The Open Archival Information System (OAIS), which is a well-known standard for digital preservation, defines the Information Package model (CCSDS, 2012). Although the information package model does not address metadata longevity very well, it provides important insights. An information package is composed of an information object and Preservation Description Information (PDI). PDI is a metadata to describe attributes required to keep information object interpretable, i.e., renderable, playable, operable, and functional in various ways. This development focuses on digital objects. A serious, overlooked challenge here is that the PDI may contain or refer metadata embedded in the object. Thus, preservation of digital objects requires long-term maintenance of metadata. This paper recognizes this challenge, and highlights the need to address metadata longevity.

This paper focuses on metadata schema longevity chiefly in the LOD environment. The remainder of the paper is organized as follows: Section 2 discusses a framework for long-term maintenance of metadata schemas from the viewpoints of LOD and DCAP developments; Section 3, defines entities included in metadata preservation in the LOD environment; Section 4 discusses entities of metadata and risks in long-term use of the entities; and Sections 5 and 6 include a discussion and conclusion respectively.

2. Metadata and Metadata Schemas in the LOD Environment

2.1. Basic Concepts and Words

This section provides definitions of key words and concepts used in this paper. *Metadata* is defined as "(structured) data about data." Metadata about a metadata is called *meta-metadata*. A *metadata schema* is an expression of definitions, including structural, syntactic and semantic features of metadata for an application. Thus, a metadata schema is metadata about the application metadata--essentially the meta-metadata for the application. Controlled vocabularies used within metadata schema are referred to as *metadata vocabularies*, and loosely classified as *property vocabularies* or *value vocabularies*. Metadata vocabularies may be defined using a formal definitions scheme such as RDF Schema and OWL in the LOD environment. The Dublin Core Metadata Element Set (DCMES) of 15 elements, is a typical property vocabulary. Subject headings such as Library of Congress Subject Headings (LCSH) and Medical Subject Heading

[1] Metadata Registries. All sites retrieved May 27, 2016
The Dublin Core Metadata Registry, from http://dcmi.kc.tsukuba.ac.jp/dcregistry/
MetaBridge, from https://www.metabridge.jp/infolib/metabridge/menu/?lang=en

(MeSH) are typical value vocabularies. A term included in a metadata vocabulary is called *metadata term*. DCMI Metadata Application Profiles (DCAP) define structural constraints for metadata and application specific requirements. The DCAPs and the DCMI's layered metadata interoperability model are key components for metadata longevity, and integral to issues in this paper.

Metadata is indispensable for searching, managing, and processing data object instances. Moreover, a metadata schema is indispensable for correctly creating and interpreting metadata. As defined above, a metadata schema is meta-metadata. In the LOD environment, metadata schemas may be shared on the Web along with metadata instances rendered with the existing schemas. As a result, the guiding metadata schemas must be interpretable by both humans and machines that comply with the LOD environment. There are schemes to define metadata schemas formally. These schemes that define metadata schemas are metadata about meta-metadata, i.e., meta-meta-metadata. Thus, this iteration of "meta" seems endless, although it is crucial to: 1) understand the levels of "meta-", 2) maintain and preserve metadata over time, and 3) keep descriptions of different levels of "meta" interpretable for machines and humans over time. Figure 1 shows relationships among "meta"-entities. In this Figure, it should be noted that all entities should be maintained for the long-term use of the object instance.

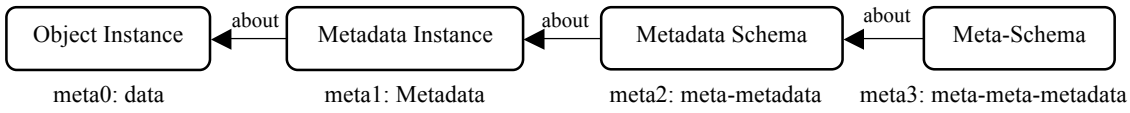


FIG.1. "meta-" relationships.

2.2. Metadata Preservation in the Linked Open Data Environment

There are various types of metadata, e.g., bibliographic descriptions, product descriptions, and rights notices. Researchers categorize these types, e.g., Lagoze (1996) and Greenberg (2005). In conventional library information systems, bibliographic records are stored in a database. Bibliographic databases are migrated from an old system to a new system many times. Metadata schemas are used for the migration, and system interoperability, and metadata schema documents need to be maintained along with the migrations. Thus, long-term maintenance of bibliographic records in a conventional system is carried out as long-term maintenance of the bibliographic database. Revision history of its bibliographic description scheme is generally recorded in its schema document when the database schema is revised.

This common maintenance practice does not, however, synchronize with the LOD metadata environment because of the fundamental difference of characteristics of metadata instances. In the LOD environment, a metadata instance is realized as an XML object which can be transferred and shared on the Web. Metadata schemas and vocabularies need to be maintained in order for metadata instances to be interpretable consistently over time. Figure 1 illustrates the requirement to maintain metadata in order to keep object instance interpretable by machine.

Machine interpretability of metadata, a key contribution of this work, is defined as follows:

- A machine driven function supporting metadata search and display, or other processes, which can automatically identify a metadata term and select a function in accordance with the meaning of the term, the metadata is full-machine interpretable.
- In the case there are revisions which do not impact machine interpretability of metadata instances but affects human interfaces, e.g., human readable labels of metadata terms, the metadata is semi-machine interpretable.

Machine interpretability of metadata means full-interpretable and/or semi-interpretable metadata.

The goal of long-term metadata maintenance is to keep a metadata instance machine interpretable over time and as intended from when the instance was created. Therefore, keeping

metadata machine interpretable consistently and over time is the primary issue in metadata longevity.

Thus, long-term maintenance of metadata instances in the LOD environment is fundamentally different from the long-term maintenance of digital objects modeled by the OAIS standard. The metadata longevity model introduced in this paper encompasses the OAIS standard for preservation of digital object instance. Instances of meta-0 level in Figure 1 are primarily included in this category. Some instances of meta-1, -2, and -3 level realized as a document for human readers should be preserved as a digital object instance. On the other hand, other instances in those levels realized as a first class object should be maintained without losing consistent machine interpretability.

2.3. Dublin Core Application Profiles and Metadata Interoperability Model

DCAP presents a generalized model of metadata schemas and their components (Heery & Patel, 2000). The Singapore Framework of DCAP shown in Figure 2 defines the components of a metadata schema for an application and related components such as metadata vocabularies. A definite separation of metadata terms and structural features is the key feature of DCAP. The Singapore Framework defines five components of an application profile – Functional Requirements, Domain Model, Description Set Profile, Usage Guidelines, and Encoding Syntax Guidelines. These components and metadata terms should be well maintained for interoperability of metadata across communities and over time.

DCMI defines a simple layered model to present levels of metadata interoperability shown in Figure 3. In the model, the lowest layer (Level 1) is interoperability given by shared informal term definitions and the highest layer (Level 4) is DSP interoperability given by shared formal vocabularies and constraints. Nagamori and Sugimoto (2004) defined a three-layered model for metadata interoperability based on the Application Profile concept shown in Figure 4.

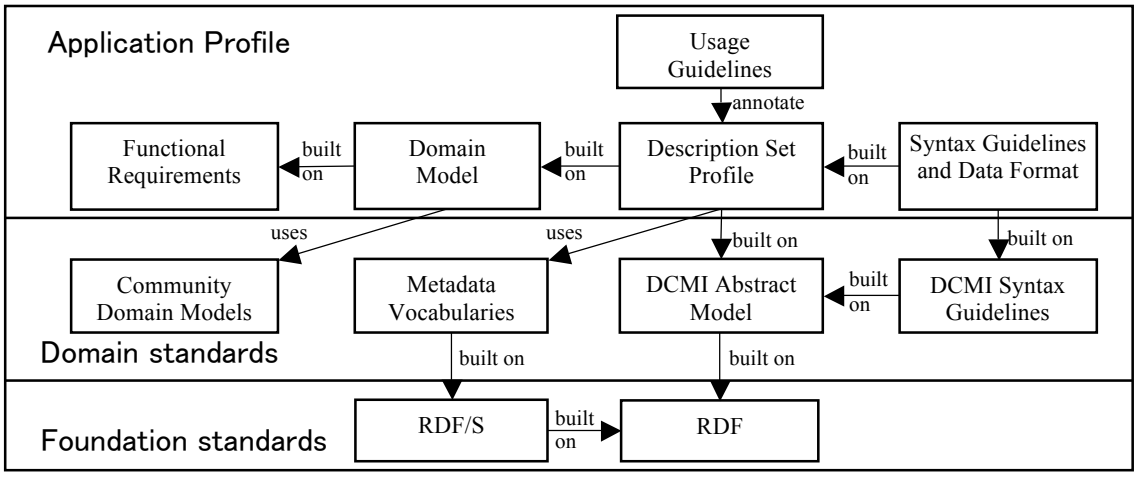


FIG.2. Singapore Framework of DCMI Application Profile (Nilsson, Baker, & Johnston, 2008).

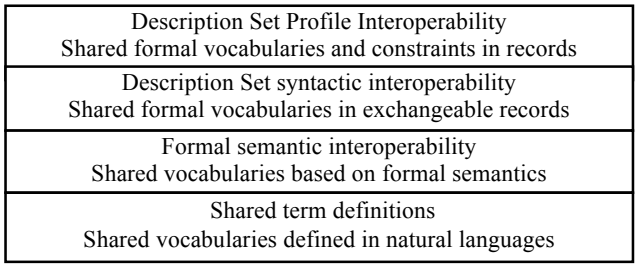


FIG.3. Interoperability Levels of DCMI (Nilsson, Baker, & Johnston, 2009).

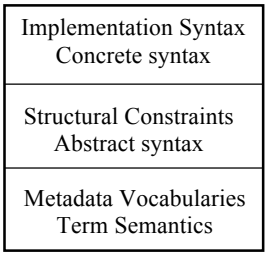


FIG.4. A Layered Model of Metadata Schema (Nagamori & Sugimoto, 2004).

A metadata schema registry, which is a repository of metadata schemas and terms, can be used as a basis for sharing metadata schemas and terms on the Web. Metadata registries are not a permanent service but have a crucial role to keep the meaning of metadata terms for metadata preservation. For instance, the Technical Registry PRONOM at the National Archives of UK collects and maintains file format information to help digital preservation.

3. Metadata Preservation as Temporal Interoperability of Metadata

3.1. Metadata Preservation Facets

The result of our work to date reveals a set of facets for the long-term maintenance of metadata – entities in different meta-levels, preservation description categories, requirements specific to metadata preservation in the LOD environment, and other aspects. Figure 5 summarizes the facets described in the paragraphs below. Versioning of metadata schemas is a crucial aspect for the long-term maintenance of metadata. We discuss versioning further, below, as a part of long-term preservation and provenance description of metadata and metadata schemas.

(1) Facet 1: Entity Format Types – Document Files, Databases, XML Encoded Texts

Longevity management of metadata entities depends on the implementation formats of entities to be preserved. For example, it may be often the case that a metadata instance is stored in a database and an XML encoded instance is created when downloading the instance from the database. In the LOD environment, any instance which should be identifiable as a resource has to be given a URI. Maintaining URIs consistent is one of the key issues for metadata permanence.

(2) Facet 2: Entity Types – Meta-Levels

As shown in Figure 1, there are instances of different meta-levels from level 0 to level 3. This paper assumes any instances of these four categories are realized in a digital form, although they may be realized as a non-digital instance, e.g., a printed document. The longevity of an Object Instance is a topic outside the scope of this immediate paper, given our focus metadata. Instances of meta-level 1, 2 and 3 may be implemented as a document-like instance, a database record, or an XML instance encoded in a metadata description standard, e.g., RDF. Metadata preservation may be done in three approaches – document preservation, database preservation and XML encoded instance preservation in accordance with requirements in each meta-level.

(3) Facet 3: Metadata Schema Components

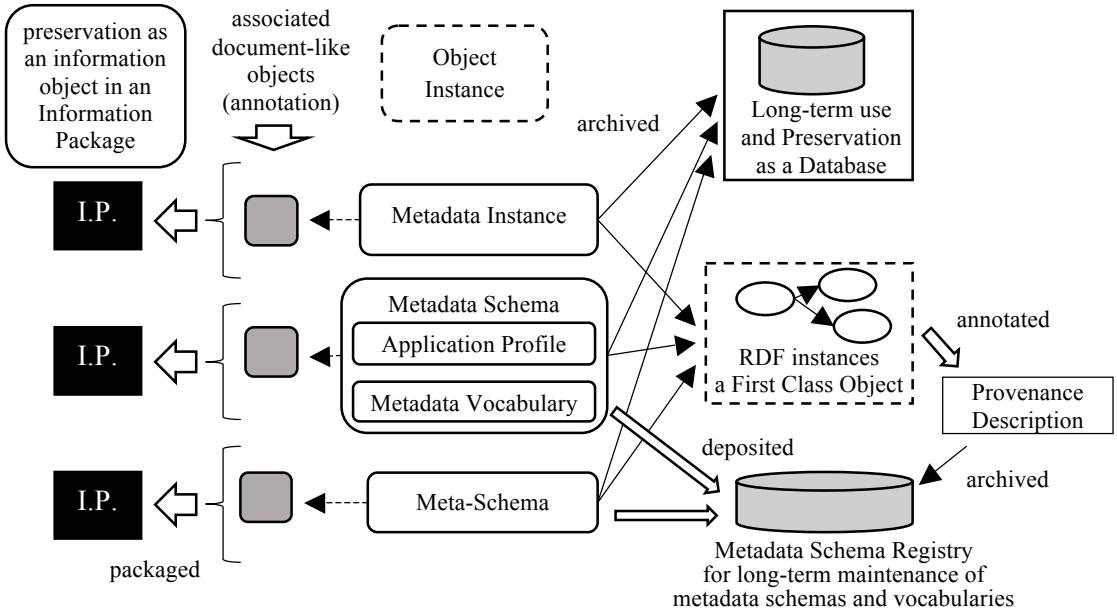


FIG.5. Metadata Entities and Preservation Options.

This facet is for metadata schema and meta-schema entities – application profiles, metadata vocabularies for certain domains and domain-neutral standards for metadata description such as XML and namespaces. For example, Description Set Profiles and Domain Models of Singapore Framework are encoded in a formal scheme and other components are expressed as natural language texts. Preservation strategy of these components depends on the entity format types.

(4) Facet 4: Dynamic Entities

Cases whereby metadata terms are removed from or added to a metadata schema, and when a new metadata schema is created by aggregating two existing schemas. In such cases, we often create a mapping table to map an old schema to a new schema. The mapping tables should be recorded as well as those schemas.

(5) Facet 5: Documentation

Any document entities and activities may be recorded for use in the future. The document entities have to be preserved as a part of metadata preservation. Contextual information, which may not be explicitly described in metadata schema entities, may be found in the documentation entities.

3.2. Related Research – Digital Preservation, Archiving and Provenance Description

Provenance is key for the maintenance of metadata. OAIS defines a package-based model for preserving digital objects. The OAIS model is applicable to any digital object of the meta-levels in the case we preserve it in a package. However, this model does not seem adequately support active entities directly accessed from other dynamic entities that change content.

Web archiving is a related area for this study. Internet Archive is a very large provider of archived Web resources. Memento defines a framework to keep old URIs consistently usable (Van de Sompel, Nelson, & Sanderson, 2013), and provides an exemplary way that may assist our work. Allowing the temporal tracking and overall path noting the history of a schema is significant for metadata longevity. This type of work can also inform URI management – one of the fundamental requirements. Keeping metadata as a Web page may be within the scope of Web archiving, but consistency management of metadata schemas is out of their scope.

In this paper, we focus on longevity of metadata entities as a first class object in the LOD environment, so that we do not focus on the longevity of metadata as a packaged object or Web page. Longevity of databases which store metadata entities is also out of the focus of this study.

Provenance descriptions track changes of metadata instances and metadata schemas. Provenance can include a series of descriptions of events for metadata instances, metadata schemas, vocabularies, and other related entities. An event description may be associated with metadata objects, e.g., agents, reasons, activities, etc. Provenance description is an important issue for metadata (Eckert, 2013). W3C has defined a provenance description model for the Web, i.e., W3C PROV (Groth & Moreau, 2013). The authors have developed a provenance description model for DSPs based on W3C PROV and RDF. This model defines Addition, Deletion and Revision activities based on W3C PROV and the DCMI Application Profile. We have experimentally applied the model to describe provenance among the versions of DPLA Metadata Application Profiles (Li, Nagamori, & Sugimoto, 2015; Li & Sugimoto, 2016).

Here, we point to metadata schema registries as services that keep metadata schemas interpretable by both machines and humans over time. There are registry services that provide metadata vocabularies and terms.[2] However, these services do not provide functions for storing a metadata schema defined for an application. DCAPs as a conceptual model for application

[2] Metadata schema and vocabulary services. All sites retrieved May 27, 2016
Linked Open Vocabularies (LOV), from http://lov.okfn.org/dataset/lov
Schema.org, from http://schema.org/docs/schemas.html
Open Metadata Registry, from http://metadataregistry.org/about.html

metadata schema are being widely recognized. However, schema registry for DCMI AP is still a nascent area of research.

4. Risks in Metadata Longevity

4.1. Aspects for Longevity Risks of Metadata

Longevity of metadata and metadata schemas reveals a number of associated risks, which are reviewed in this section. What is risk for metadata longevity? Where do longevity risks exist? How can we find the risks? These are fundamental research questions for this study. The following sentences are typical risks:

Risk-A: Metadata schema of this metadata instance is unknown.
Risk-B: We cannot display a metadata creation guideline documents correctly.
Risk-C: An application uses a standardized metadata vocabulary but the name of the vocabulary is unknown.
Risk-D: Definition of a metadata term is not identifiable by URI given to the term.

These risks occur for many different reasons, e.g., insufficient documentation, insufficient information transfer when downloading metadata, inappropriate maintenance of documents, non-persistent URI, etc. We focus on the risks from three aspects in the following sections, (1) metadata instances, (2) application-based schemes, and (3) shared vocabularies and terms. The Singapore Framework, RDF and URI are the underlying framework for the discussion in this paper. As URI is the base identification scheme of any instances, persistency of URI is a fundamental requirement for the longevity of metadata. We discuss this issue later in this paper.

4.2. Metadata Instances

Temporal interpretability of metadata instances depends on the availability of the metadata schemas for those instances over time. Metadata instances can be classified into two classes – primary and secondary metadata. The primary metadata is a metadata instance stored in a metadata database, or embedded in a source resource such as books and Web pages. Non-primary metadata instances are secondary metadata, which may be created by copying and editing primary metadata. Provenance information of the secondary metadata is crucial in order to keep its consistency but is not always created.

When a metadata database is created by re-organizing downloaded metadata, the metadata instances become the primary data, with the underlying scheme guiding the new metadata database organization. In this case, provenance description should be included as contextual information of the newly created database, e.g., information about the source resource and schema.

Longevity risks of metadata instances depend on whether a metadata is primary or secondary. Longevity risks of primary metadata instances exist in the management of their metadata schemas. However, in the case of secondary metadata, their longevity risks are not only in the same factors but also in keeping provenance information of the metadata instances consistent. In the reality, provenance information may not be recorded in the most cases of copying metadata. This means that there is no way to keep those metadata consistent over time.

4.3. Application Profiles – Structural Constraints and Syntax

Long-term maintenance of metadata schemas is crucial for metadata longevity, a key point in this paper. The Singapore Framework (SF) offers support here, with five components – Functional Requirements, Domain Model, Description Set Profile, Usage Guidelines, and Encoding Syntax Guidelines. Although many existing metadata schemas do not specifically define each SF conformant application profile component, many schemas have these aspects integrated. This paper uses SF as the basis for discussion because it clearly states the aspects that should be included in the definition of a metadata schema for any application. SF explicitly

defines dependency among the five components of SF and relationships between DSP and metadata vocabularies. This clear separation of the metadata schema components helps the maintenance of metadata schemas. From this viewpoint, we can reduce risks of metadata longevity by using SF for metadata schema definition.

Maintaining these five components depends on the formality of their descriptions. Natural language text documents should be maintained as a textual file, formal descriptions should be maintained in accordance with the formal schemes of the description, i.e., UML and RDF. Metadata schema instances presented in RDF may be stored as a set of triples, which means that we would need to maintain them as a set of metadata instances but not as a textual document.

4.4. Metadata Vocabularies and Terms

SF defines metadata vocabularies in a layer beneath the application profiles. Metadata vocabularies defined for a domain but neutral to particular applications are defined in that layer. This metadata vocabularies layer is defined above a layer in which domain neutral constructs for implementing metadata are defined, e.g., RDF, XML and other Internet standards. This separation is the fundamental issue from the viewpoint of metadata interoperability, meaning not only interoperability across communities, but also interoperability over time. URI, RDF and OWL are the standard schemes to define the terms in the LOD environment. Definitions of the terms and vocabularies in these schemes may be presented in a document, stored in a database, or realized as a first class object encoded in XML/RDF. From the viewpoint of machine interpretability of metadata, keeping the database or the first class object accessible is a key. From the viewpoint of human readability, any form of these forms is acceptable.

Versioning information for both terms and vocabularies is key for supporting long-term maintenance. There are different cases of versioning – versioning of DCMI terms is term-basis but versioning of decimal classifications is vocabulary-basis. As precise meaning of a metadata term depends on its versions, keeping version information is crucial for long-term use of metadata.

Maintenance authorities for metadata standards are supposed to be stable but may disappear over time. There are metadata schema registries and Web sites, which provide definitions of metadata vocabularies and terms. Multiple copies of descriptions are a double-edged sword – on one hand multiple copies are robust for keeping the content safe for the future, but on the other hand, multiple copies may cause troubles in consistent maintenance of versions.

4.5. Other Factors

Persistency of identifiers: URI, commonly used to identify Web and Internet resources, is used to identify metadata schema instances, metadata terms, and vocabularies in the LOD environment. Metadata term URIs are sometimes used, further, for term definitions, and making them accessible by the URI, i.e., resolvability of URI. Persistency of URL is the fundamental requirement for the longevity of metadata terms and vocabularies. Resolvability of URI is not mandated for the persistency of URI.

Metadata mapping tables: Metadata mapping tables are often created for many purposes – merging two or more metadata datasets, metadata harvesting, federated search, and so forth. Metadata mapping table is a crucial resource in the long time line of metadata maintenance. As a metadata mapping table is a kind of metadata, i.e., description of metadata mapping, we can apply the model proposed in this paper for the metadata mapping table.

Contexts: Contextual information of metadata and metadata schemas is crucial for their longevity. However, it is hard to describe the contextual information perfectly. Every metadata schema designed for an application has its contexts that may or may not be described as a part of the schema. For instance, descriptions about selection process of metadata terms from standard vocabularies are crucial to know the context of the metadata schema and to correctly interpret

metadata. In theory, it is feasible to keep such description because SF includes components that may include contextual information. However, those descriptions tend to be not provided

5. Discussions – Provenance Description for Risk Management for Metadata Longevity

Over the last few years there have been national and international calls targeting the archiving and preservation of research data. National and global agencies require data deposition and they require researchers with funding to make their research data accessible, and reusable. In connection with this significant development, it seems equally if not more important to call for the publication, preservation, and archiving of metadata standards, and the levels and facets reviewed above, to support long-term interpretability.

Metadata schema documentation is required for proper maintenance of metadata instances. We learned that not many LOD datasets provide information about their metadata schema (Honma, Tanaka, Nagamori, & Sugimoto, 2014). Proper versioning information of metadata schemas is necessary but application schemas tend to loose consistent maintenance of the information. These sorts of information should be maintained in a machine interpretable form rather than a human readable form from the viewpoint of keeping metadata machine interpretable over time.

Formal description of provenance of metadata schemas is essential to cope with this maintenance problem. The authors have developed a provenance description model based on DCMI Description Set Profiles and W3C PROV in order to formally express provenance of description set profiles in RDF and use the description for automated consistency checking.

Another crucial issue is to use metadata schema registries for long-term maintenance of metadata schemas. Current metadata schema registries and related services provide current information about metadata vocabularies. MetaBridge provides a function to store/provide description set profiles in RDF but it has only a simple versioning function to replace an old version by a new version. Our study on provenance description of DSP shows that RDF-based provenance description helps maintenance of metadata schemas. It is necessary to be able to identify a version of a metadata term used in a DSP. A DSP is linked to a metadata term by URI of the term. However, as URI does not convey any version information of the term, we need to use provenance information of the DSP to maintain the linkage between the DSP and its corresponding version of the metadata term, which may be implemented in a metadata schema registry for metadata schema preservation.

URI is not persistent but metadata terms have to be consistently identifiable over time in the LOD environment. Thus, persistency of URI is essential for long-term maintenance of metadata schemas. Persistent URIs rely on persistent URI resolvers. Metadata schema may be able to keep the definition of a metadata term associated with its URI. LOCKSS metaphor may work for keeping definitions of metadata terms and application profiles consistent. We need collaborating metadata registries for keeping metadata schemas safe over time.

6. Concluding Remarks

In this paper, we have discussed risks in metadata longevity by analyzing various entities of metadata from different aspects. We have proposed to use provenance description and metadata schema registry for the risk management in this paper. Preservation of digital objects is a well-known research topic for digital curation and archiving. Conventional digital preservation is oriented to preservation of a primary entity such as documents, games, pictures, etc. Metadata preservation has been discussed within the scope of conventional digital preservation. However, in the LOD environment, there are many new issues for long-term and consistent use of metadata. Conventional OAIS-based preservation is a frozen preservation because we need to retrieve and open information packages. On the other hand, in the LOD environment, we will have many metadata instances stored in our files as a first class object. Keeping these instances consistently interpretable is crucial in such an environment, which may be called unfrozen archive. The

development of the concept of DCAP has contributed to clarify requirements for metadata interoperability. However, temporal interoperability of metadata is still not well studied yet.

It is widely known that term definitions and term usage changes over time, and can further change due to domain use. In this paper, we mentioned this issue as contexts. We understand that it is important to include the context explicitly in the metadata schema management process but it is challenging to explicitly and consistently describe the contexts based on the underlying data model of LOD. Management of contextual information for the longevity of metadata and metadata schemas is a fundamental issue but is left for our future study.

Acknowledgements

This work was supported in part by JSPS Kaken Grant-in-Aid for Scientific Research (A) #16H01754 and #25240012.

References

CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS). Retrieved May 27, 2016, from http://public.ccsds.org/publications/archive/650x0m2.pdf.

Eckert, Kai. (2013). Provenance and Annotations for Linked Data. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2013, 9-18. Retrieved May 27, 2016, from http://dcpapers.dublincore.org/pubs/article/viewFile/3669/1892.

Greenberg, Jane. (2005). Understanding metadata and metadata schemes. Cataloging & classification quarterly, 40(3-4), 17-36.

Groth, Paul, and Luc Moreau (Eds.). (2013). PROV-Overview. Retrieved May 27, 2016, from https://www.w3.org/TR/prov-overview/.

Heery, Rachel, and Manjula Patel. (2000). Application Profiles: Mixing and Matching Metadata Schemas. Ariadne, 25. Retrieved May 27, 2016, from http://www.ariadne.ac.uk/issue25/app-profiles.

Honma, Tsunagu, Kei Tanaka, Mitsuharu Nagamori, and Shigeo Sugimoto. (2014). Extracting Description Set Profiles from RDF Datasets using Metadata Instances and SPARQL Queries. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2014, 109-118. Retrieved May 27, 2016, from http://dcpapers.dublincore.org/pubs/article/view/3706.

Lagoze, Carl. (1996). The Warwick Framework: A container architecture for aggregating sets of metadata. D-Lib Magazine, July/August,1996. Retrieved May 27, 2016, from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

Li, Chunqiu, Mitsuharu Nagamori, and Shigeo Sugimoto. (2015). Temporal Interoperability of Metadata: an Interoperability-based View for Longevity of Metadata. Proceedings of the 6th International Conference on Asia-Pacific Library and Information Education and Practice, 2015, 199-209.

Li, Chunqiu, and Shigeo Sugimoto. (2016). Provenance description of metadata for long-term maintenance of metadata application profiles, unpublished draft.

Nagamori, Mitsuharu, and Shigeo Sugimoto. (2004). A Metadata schema framework for functional extension of metadata schema registry. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004. Retrieved May 27, 2016, from http://dcpapers.dublincore.org/pubs/article/viewFile/764/760.

Nagamori, Mitsuharu, Thomas Baker, Tetsuo Sakaguchi, Shigeo Sugimoto, and Koichi Tabata. (2001). A Multilingual Metadata Schema Registry Based on RDF Schema.Proceedings of the International Conference on Dublin Core and Metadata Applications, 2001, 209-212. Retrieved May 27, 2016, from http://dcpapers.dublincore.org/pubs/article/view/660.

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. Retrieved May 27, 2016, from http://dcpapers.dublincore.org/pubs/article/view/3632.

Nilson, Mikael, Thomas Baker, and Pete Johnston. (2008). The Singapore Framework for Dublin Core Application Profiles.Retrieved May 27, 2016, from http://dublincore.org/documents/2008/01/14/singapore-framework/.

Nilson, Mikael, Thomas Baker, and Pete Johnston. (2009). Interoperability Levels for Dublin Core Metadata. Retrieved May 27, 2016, from http://dublincore.org/documents/2009/05/01/interoperability-levels/.

Van de Sompel, Herbert, Michael L. Nelson, and Robert Sanderson. (2013). HTTP Framework for Time-Based Access to Resource States-Memento. No. RFC 7089 (Status: Informational). Retrieved August 31, 2016, from https://tools.ietf.org/html/rfc7089.

Metadata Profiles

Towards the Development of a Metadata Model for a Digital Cultural Heritage Collection with Focus on Provenance Information

Susanne Al-Eryani
State and University Library Göttingen
salerya@sub.uni-goettingen.de

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

Abstract

This project report describes the first steps of the development of a metadata model for the contextualization of heterogeneous objects from different cultural heritage collections with focus on provenance information. The project started with the assumption that aims and objectives of researchers working with cultural heritage collections differ from discipline to discipline. Accordingly, use cases and requirements for the description of objects are heterogeneous. To provide a model that would be usable not only within but also across academic disciplines the project needed to know where these requirements differ and where they match. Therefore the first part of the project was focused on the investigation of use cases and requirements. On the base of the common requirements a generic model will be build that allows the merging of data from a variety of disciplines using different metadata standards. The model's structure will be a combination of prevalent metadata standards mapped to each other. Another peculiarity of the model will be the modular design of micro-ontologies, sets of domain-specific class structures that are, nevertheless, available on a meta-level in terms of substructures. Applying the DCMI dumb-down principle these subproperties and subclasses will be assigned to a who-what-where-when model, a base structure for the description of objects.

The project divided the work process of the project into seven steps. As the project is still work in progress, only four steps will be explained in detail in this report. The three remaining steps will be presented in an outlook.

Keywords: metadata model; metadata standard; digital cultural heritage collection; provenance information

1. Introduction

The predicted shift or extension from the concepts of the current World Wide Web, the so called Web 2.0, to those of Web 3.0 is in full swing and the debates on how to establish an appropriate base in the context of this challenge often end up with a big question mark. The expression 'Web 2.0', coined by Darcy DiNucci (1999) and made popular by Tim O'Reilly (2005) at the Web 2.0 Conference, held in San Francisco in 2004, stands for an interactive medium that can be described as a web of documents connected by hyperlinks suitable for human consumption. In contrast, Web 3.0, also known as 'Semantic Web', is a web of structured data conveying semantic meaning and connected by semantically meaningful links. This web of machine-readable data will support people's needs, for example, to create data stores on the Web or to achieve precise information from an unmanageable number of options (see W3C, 2015). But, is there a golden road that would lead to a satisfying supply of information in the net in order to make data accessible and searchable according to most diverse requirements, and in addition, that would provide a base for embedding data appropriately into a semantic net of information? Or must metadata specialists and information professionals working in cultural heritage institutions develop their very special metadata model for "their own data" within "their own institution", when preparing the data for future requirements? However, the application of different standards leads to enormous challenges when it comes to the interlinking and automatic

processing of data. The overall aim should be an extraction of the main concepts from the wide range of knowledge fields, transferred into a modest quantity of metadata schemes that would be sustainable and usable within different professional contexts. In a project, the work of which will be presented in this paper, the attempt is being made to create a metadata model that would meet this requirement.

The three-year project with its somewhat cumbersome designation Developing interoperable metadata standards for contextualizing heterogeneous objects, exemplified by objects of the provenance von Asch (short ASCH; see http://asch.wiki.gwdg.de) at the State and University Library of Göttingen (SUB), Germany, is lead by the SUB and the Institute of Social and Cultural Anthropology, in collaboration with the Metadata Group and the department Digital Library of the SUB and several collections of the Göttingen University (named in section 2), and is supported by the Deutsche Forschungsgemeinschaft (DFG). As the name implies, the project's work focuses on the development of a metadata model, which means the integration of various interoperable metadata standards (i.e. metadata schemes or element sets and corresponding application profiles) for the contextualization of heterogeneous objects of cultural heritage collections.

By describing resources of digital collections, the use of metadata standards and authority data is an essential technical precondition for making them identifiable and retrievable in the net. Considering the existing variety of information and the diverse but mostly not semantically defined web of relationships that link information to other information, the creation of a universally valid data model would be desirable but seems to be unconceivable, at least until present. Nevertheless, why should it not be possible to develop a very generic model for single segments of shared cultural knowledge that would be extensible in accordance to the requirements of individual research disciplines established in the various cultural heritage and scientific institutions? Would it be a realistic aim to generate a basic template that would be reusable and extendable in different contexts? Tim Berners-Lee formulates four assumptions for the interconnectedness of data: (1) things have to be named by URIs (Uniform Resource Identifier), (2) the URIs should be HTTP URIs, (3) useful information should be given on these URIs by using certain standards (RDF[1], SPARQL[2]), and (4) links to other URIs should be included (Berners-Lee, 2006). This linkable data, known as Linked Data, can be understood as a "set of best practices for publishing and connecting structured data on the Web using international standards of the World Wide Web Consortium" (Wood et al., 2014).

In order to provide linkable data, the resource descriptions must correlate to common metadata standards. A fast increasing number of scientific institutions, archives, libraries and museums are eager to prepare and edit their databases in order to make their digital resources interoperable even across institutional borders. This challenging task presupposes the application of appropriate metadata standards. There are a number of standards that fulfill the requirements of the different scientific and cultural heritage institutions: libraries are using, for example, MARC 21 and MODS (Metadata Object Description Schema), the application of EAD (Encoded Archival Description) and EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families) is commonly used by archives, and LIDO (Lightweight Information Describing Objects) is a widespread scheme applied by museums. In the world of natural science the ABCD (Access to Biological Collection Databases) and its extension ABCDEFG (ABCD Extended for Geosciences) are a first step to provide data across institutions and disciplines as is the Darwin Core standard. In addition to the application of metadata schemes, the use of authority data becomes indispensable for the description of resources because the semantic assignment via URIs facilitates an unambiguously identification of applied terms. Examples are the LCSH (Library of Congress Subject Headings), VIAF (Virtual International Authority File) and GND (Gemeinsame

[1] http://www.w3.org/RDF/
[2] https://www.w3.org/TR/rdf-sparql-query/

Normdatei)[3] used by libraries, the Getty Vocabularies[4] providing structured terminology for different cultural fields, and a wide array of taxonomies used in natural science.

The illustration of provenance information will be of special interest by developing the ASCH model. Therefore, common standards describing aspects of provenance will be considered. A variety of ways can be found for this description of resources, because different subject areas focus different aspects by documenting the life history of objects. T-PRO (Thesaurus der Provenienzbegriffe),[5] for example, is a thesaurus to describe terms of provenance in an object-orientated manner and is used by German libraries. For an event-based description of objects, LIDO is an appropriate format mainly used by museums. CIDOC CRM (CIDOC Conceptual Reference Model) enables to illustrate provenance information with the additional option of embedding evidences to the given facts, and an abstract level for description is possible by applying the PROV-DM (PROV Data Model)[6] provided by W3C.

The ASCH model is expected to merge different metadata standards commonly used by various cultural heritage institutions on a meta-level in order to make the metadata reusable in an interdisciplinary context as done by the DDB (Deutsche Digitale Bibliothek)[7] and Europeana,[8] for example. The who-what-where-when model developed by the DDB, and the Europeana Data Model (EDM) developed by Europeana allow specific object- and event-orientated resource descriptions, but provenance information cannot be illustrated in greater depth and an explicit interlinking to external evidence is not possible. To bridge the gap between these description frames is the aim of the ASCH project. The functionality of the ASCH model will be tested by using descriptions of digitized objects compiled from certain collections that are relevant for a chosen specific provenance context. The historical background of these collections will be depicted in the following section. Afterwards the methodology of the project's work will be explained in more detail.

2. Historical Background of the Collections

Seven collections of the Göttingen University are known to house or at least to have housed objects that were sent from Saint Petersburg in the second half of the eighteenth and the beginning of the nineteenth centuries. These collections are:

- the Historic Printed Collections, Manuscripts and Rare Books at the Göttingen State and University Library;
- the Ethnographic Collection at the Institute of Social and Cultural Anthropology;
- the Skull Collection at the Department of Anatomy and Embryology, Centre for Anatomy, University Medical Centre Göttingen;
- the Historical Collections at the Geoscience Centre;
- the Coin Cabinet at the Department of Archaeology;
- the Art Collection at the Department of Art History; and
- the Museum of Zoology.

The objects of these collections share a uniting circumstance in their life history because their provenance can be traced to a certain collector who had given them to a certain institution during a certain period of time. But, the characteristics of these objects are very distinctive and therefore they became prime candidates for the development of our metadata model. The objects' history leads us to the collector Georg Thomas von Asch (1729-1807), a Russian physician who had

[3] http://www.dnb.de/gnd
[4] http://www.getty.edu/research/tools/vocabularies/
[5] http://provenienz.gbv.de/T-PRO_Thesaurus_der_Provenienzbegriffe
[6] https://www.w3.org/TR/prov-dm/
[7] https://www.deutsche-digitale-bibliothek.de/
[8] http://pro.europeana.eu/

conducted his medical studies in Germany and had received his Doctorate of Medicine at the Georg August University in Göttingen. After his return to Russia, Baron von Asch had kept up close ties to Christian Gottlob Heyne (1729-1812), the director of the Göttingen University Library. The baron was also well acquainted with Johann Friedrich Blumenbach (1752-1840), the director of the Royal Academic Museum. Between 1771 and 1806, von Asch had sent more than 120 parcels and boxes to Göttingen, filled with natural and man-made objects of a wide range in order to be incorporated into the holdings of the University Library or the Academic Museum, respectively. In the second half of the nineteenth century, the Royal Academic Museum was dissolved and its collections were distributed among the new founded departments of the Göttingen University, named above, where they are partly to be found until present (for further readings see Hauser-Schäublin; Krüger [eds.], 2007).

In many cases, the origin of the ethnographica, botanica, zoologica, coins, rocks and other natural objects, skulls, prints, manuscripts, maps and books can be traced with great accuracy. Contemporary inventory books in the collection's archives, letters, especially the correspondence between Baron von Asch and Heyne, inventory lists enclosed to the parcels and boxes, and additional object descriptions, sometimes written on wrapping paper added by the donator, shed some light on the objects' biographies. In some cases, the provenance information is incomplete, e.g. object labels went lost during a flood, and some objects were given away in exchange for other objects so that their belonging to the former interdisciplinary collection cannot be proved. In other cases, provenance information on objects is available, but the current location of the items is unknown. Via preserved evidence it might be possible to reconstruct the objects' "journeys" and to bring them back virtually to their "home collections".

3. Work Methodology

Cultural and scientific heritage institutions have found their special way to manage collections by storing objects as well as information about objects. Analog formats for recording such as inventories, card catalogues, handwritten lists, vertical files and file labels can be found in the institutions' archives, but even in front of the doors of those buildings sometimes referred to as being old fashioned and dusty the technical revolution has not stopped. Meanwhile, analog recordings mostly have been transferred into a digitized format and stored objects have been photographed and digitized. Nowadays, the digitized metadata can increasingly be found in digital information systems that allow users access either open or locally restricted (Gilliland, 2008).

Turning to our purpose, what would be best practice to develop a single metadata model encompassing data received from different institutions with different research fields that handle their resource descriptions in various ways? In which manner could provenance information as well as external evidence referring to collection objects be linked? Our work methodology to achieve a solution can be reflected in the following seven steps which will be explained below:

1. empirical survey, analysis and evaluation of gathered information;
2. formulating of use cases;
3. analysis of requirements;
4. identification of classes and relations between classes;
5. identification of properties;
6. development of application profiles; and
7. testing the model's functionality.

3.1. Step One: Empirical Survey, Analysis and Evaluation of Gathered Information

Although we were eager to reuse widespread metadata standards and not to reinvent the wheel, we abandoned applying the complete element sets provided by these standards. Instead, we carried out an empirical survey in order to take the needs and requirements of various scholarly

communities into account. Therefore, we conducted a two-day international workshop for which we invited about forty experts representing different scholarly disciplines (Anatomy, Archaeology, Computer Science, Geology, Geosciences, History, History of Art, Librarianship, Medicine, Mineralogy, Musicology, Social and Cultural Anthropology, Philology, and Zoology) who are known to be engaged in the subject matter of provenance. In small but heterogeneous focus groups as well as in the plenum we discussed questions and scenarios covering the following subjects: understandings of the term 'provenance', experiences with data bases and data exchange, use of authority data, handling of evidence proving an object's circle of live, practices concerning gathering and recording provenance information, reusability and editing of data, use of metadata standards and research infrastructures, and best practices, bad experiences and visions concerning work routines and research conditions. This form of information collection enabled us to obtain expert knowledge from representatives of the natural science and the humanities, and from archives, libraries, and museums as the three different kinds of cultural and scientific heritage institutions at once. An elaborated documentation of performance and topics of the workshop and detailed information on the round table discussions and the result analysis is available via the project's wiki (see http://asch.wiki.gwdg.de/index.php/Workshop_2015).

In addition to this workshop, about twenty one-on-one interviews with colleagues from institutions involved in provenance research or metadata creation gave us a greater understanding of their experience and wishes concerning provenance description.

Parallel to the empirical survey we analyzed the data provided by those Göttingen University collections in which donations of Baron von Asch are preserved. As it turned out (and we had expected), the data formats are as diverse as their providers and range from spreadsheets to proprietary database management systems. Furthermore, a variety of vocabularies, home-made thesauri as well as authority files, are used for the description of resources. Formatted for human consumption, a large part of the resource descriptions examined in the collections is not in an appropriate condition to be accessible and reusable with regard to a semantic interlinking on the Web. A transformation of the stored knowledge into sharable data available for a wider audience would be only feasible by structuring the data compliant to appropriate metadata standards. Therefore, the key elements of the model, its classes, properties and the relations between the entities, described allusively in the existing data, had to be identified.

3.2. Step Two: Formulating of Use Cases

The results of the workshop showed the heterogeneity of entities and relations needed for describing resources in the different disciplines, but also the similarities. It became apparent that each discipline needs its specific metadata scheme to describe their objects but that some components of the description should be reusable in other disciplines. Especially the reuse and interlinking of provenance information was seen as a step forward because it allows the contextualization of objects examined in different disciplines but once present at the same event (e.g. when a book and a tool were bought at the same time at the same place by the same person). Another aspect broadly discussed in the focus groups was how to verify the reliability of statements by metadata descriptions or interlinking with evidence.

The results were affirmed by the work and research experiences of the experts we interviewed. We clustered the reports into dimensions of topics, needs, entities, and relations and anonymized, generalized or specified the statements. Using this material we formulated case studies taking into account the recommendations of the cultural heritage aggregators DDB[9] and Europeana[10]. The short hypothetical stories of the case studies reflect research activities, describing the usage of

[9] https://pro.deutsche-digitale-bibliothek.de/teilnahmekriterien

[10] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.3_042016.pdf

metadata by users in general as well as in different contexts. These case studies were differentiated into scenarios, smaller units describing a specific usage of metadata by a user. We focused on those scenarios that many of the disciplines had in common, e.g. (1) "a user needs information about all events in the lifecycle of an item during a certain time span" (Scenario 22) or (2) "a user needs information about items that were present during an event" (Scenario 23). One or more scenarios describing individually the actions and aims of an actor were specified in each use case. Requirements were gathered from these scenarios describing the rules and constraints necessary for the realization of an action. The following use cases relevant to further work were figured out:

- **Information about resources,** described by scenarios related to the search and finding of resources on the Web in general;
- **Identification of resources,** described by scenarios related to the identification of resources;
- **Information about the history or lifecycle of resources,** described by scenarios related to information about events or activities the items of a collection were involved;
- **Information about change of use and reception of resources,** described by scenarios related to the change of use or reception an item underwent in its history;
- **Proof of information by evidence,** described by scenarios related to the description or search of evidence that proof the reliability of information;
- **Reliability of statements**, described by scenarios related to information about the description of statements by statements;
- **Access to resources**, described by scenarios concerning the usability of resources; and
- **Reuse of data,** described by scenarios related to the use of metadata descriptions by others (for more information see http://asch.wiki.gwdg.de/index.php/Use_Cases).

3.3. Step Three: Analysis of Requirements

The use cases, consisting of one or several scenarios, helped to achieve an abstract level by analyzing the possible interactions between an actor and a system. The fundamental structure of a scenario is composed of an identified actor and one or several identified goals this actor is pursuing. With the object to gain an identified goal, one or more requirements can be necessary or even be mandatory. E.g. the above-mentioned Scenario 22 is connected to two requirements: (1) Requirement 20 (Item descriptions must be interlinked with 1-n events in the lifecycle of the item) and (2) Requirement 27 (An event in the lifecycle of an item must be related to 0-n date information). In order to organize the gathered material, we subdivided the requirements according to three aspects:

- Requirements concerning the end-user: One of the determining factors for modeling a scheme is the context of usage it shall apply to. Therefore, it is indispensable to examine the field of use, the target group, and the language of data that would be appropriate.
- Requirements concerning the metadata: It has to be analyzed which properties of entities and what relationships between these entities must be taken into consideration.
- Requirements concerning the system: The functional settings of the system have to be examined because they are responsible for the accessibility to the data as well as for its representation and retrieval.

All in all, we figured out about eighty requirements.

As we defined the class Resource to be the superclass of all classes used in the ASCH model, the requirements concerning this class would be valid for all subclasses. More detailed and specified requirements were additionally assigned to the subclasses. Within the framework of this paper we can list some examples only (1) for superclass: e.g. resource descriptions must be machine readable, resource descriptions must be compliant to the one-to-one principle, resources

must be interlinked with each other using unique, machine readable and persistent identifiers; (2) for the subclass Event, for example: e.g. an event in the lifecycle of an item must be related to 1-n items, an event in the lifecycle of an item must be related to 0-n places, an event in the lifecycle of an item must be related to 0-n date information. A detailed documentation is to be found in the ASCH wiki (see http://asch.wiki.gwdg.de/index.php/Use_Cases).

Concerning the provision of data, yet another aspect is significant – according to the Semantic Web and Linked Data, an application should not only be restricted to the concrete requirements of individual end-users, it also should keep a close eye on the possibilities of the networking opportunities given within the WWW. It is precisely for this reason that requirements, resulting from "metadata standards" used by the target communities, are taken into consideration when a certain profile shall be developed (see Zeng; Qin, 2008). Therefore, classes used in the ASCH model were aligned to classes from metadata schemes commonly used in the cultural heritage world.

3.4. Step Four: Identification of Classes and Relations between Classes

According to the requirements elicited from the use cases and scenarios, and considering the range of classes applied in common metadata schemes relevant for the description of collection items and their provenance, following classes (i.e. the superclass Resource and twelve subclasses) were identified to be used in the ASCH model:

- **Resource:** The superclass of all classes used in the model, all requirements valid for this class are also valid for all other classes of the model.
- **Metadata set:** The machine-readable description of a single resource represented by statements.
- **Item:** A real world thing in a collection.
- **Evidence:** A resource proving the reliability of a statement about a resource.
- **Event:** An activity in the lifecycle of a resource.
- **Time:** A time-span related to a resource via an activity or as a topic.
- **Agent:** A person, organization or group related to a resource via an activity or as a topic.
- **Place:** A geographic location related to a resource via an activity or as a topic.
- **Digital representation:** A digital resource depicting an item.
- **Collection:** An aggregation of items.
- **Statement:** A predication about an item.
- **Holding:** The place an item is located.
- **Concept:** A term from an authority used as a value in a resource description.

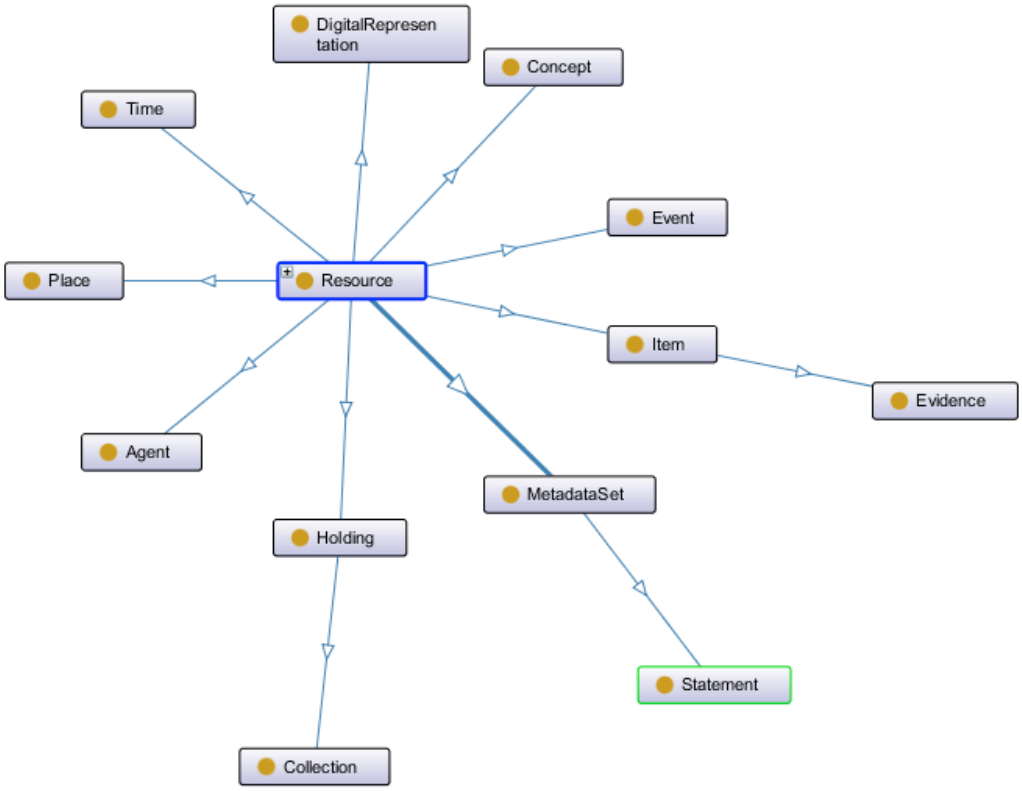


FIG. 1. The ASCH model classes and relations.

As the project focuses on the standardized description of the semantic contextualization of objects and especially provenance information about these objects, we considered only those standards relevant that are appropriate for a semantic contextualization. At present, these are Dublin Core (with DC Metadata Element Set and DCMI Metadata Terms), EDM, PROV-DM, CIDOC CRM, and Darwin Core (DwC), an extension of DC for biodiversity information. Scope of the selection of these standards is interoperability and the cross-domain use of metadata. In the DCMI Glossary[11] the term 'interoperability' is defined as "the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner." Developed for different scientific and institutional fields, these standards are focusing on diverging requirements. DCMI e.g. provides standards especially for a generic description of resources on the Web, but is also a Linked Data compliant standard. PROV is developed as an RDF standard for the description of provenance information of web resources, leaving a further description of the resource to other standards. CIDOC CRM, a semantic model that forms a base for other metadata standards (e.g. LIDO or EDM), is concentrating on the events in the lifecycle of items and DwC allows the detailed taxonomical assignment of items. RDF as one requirement to make data linkable will be used with terms of the chosen standards and evidence shall be interlinked with object descriptions because one of the requirements relevant for the research community is the interlinking of metadata descriptions of objects with parts of text in evidence

[11] http://www.dublincore.org/documents/usageguide/glossary.shtml

encoded in TEI[12] and referencing to these objects. In this context it will be possible to describe who made which statement when and where, and how reliable a statement is.

Figure 2 illustrates the provenance component of the ASCH model. All classes in the ASCH model will be identified as subclasses of the above listed RDF compliant ontologies. So an item of a zoological collection may be described using DwC and thereby be compliant with other data from the same discipline. Then the provenance description via the Event class is using DwC properties and classes parallel to PROV properties, and classes where the PROV properties and classes are a hub that allows the contextualization of this data with data from other disciplines describing the same event.

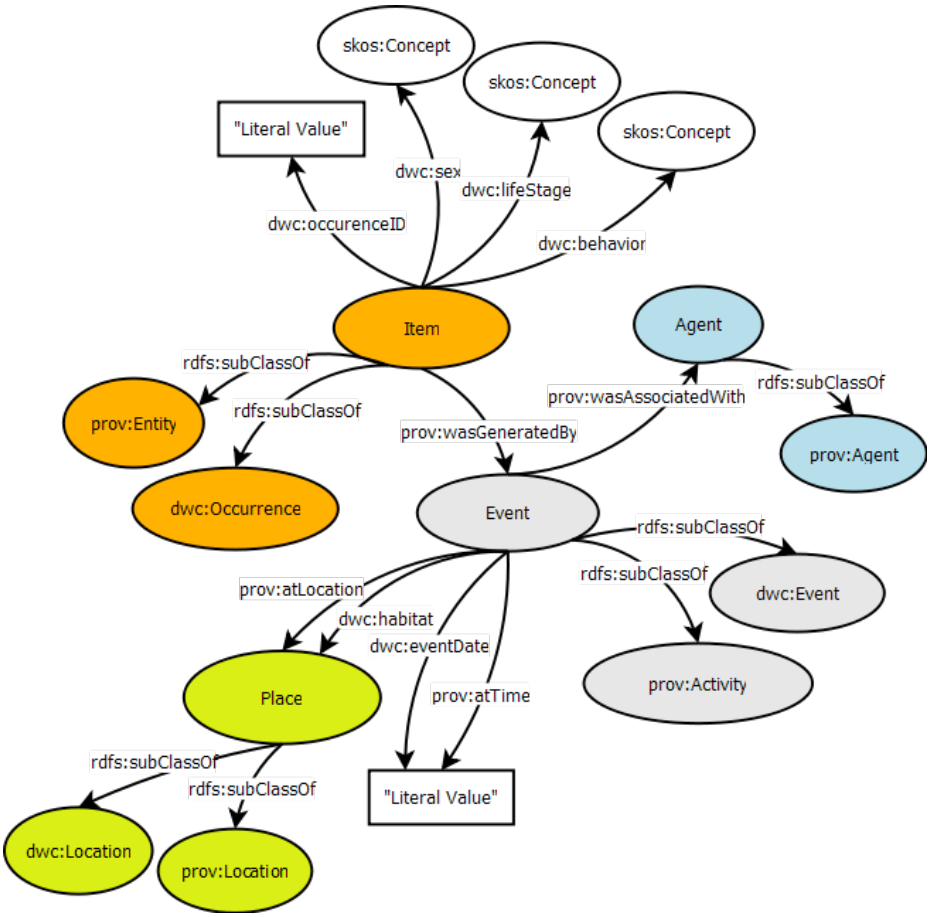


FIG 2: Description of zoological items using DwC and PROV Ontology

4. Conclusions and Future Work

To help making the huge amount of digitized cultural heritage objects accessible, it is indispensable to align metadata about these objects using a hub for those components of the description that are relevant across disciplines. This would clear the way to the interoperability of data across the borders of the various disciplines because it allows the use of domain specific metadata terms where necessary and common used terms where possible. To find out which components are usable for such a hub, we discussed the differences and similarities with experts working in natural science and humanities and in different cultural heritage institutions. The

[12] http://www.tei-c.org/

result was an abundance of information allowing us to identify those scenarios and requirements relevant for all experts independent from their background and research environment. Based on these results we started to identify the relevant classes and aligned them to RDF compliant metadata schemes used in the different domains. With the definition and alignment of these classes we finished the first four steps of our project.

Step five of the project's working plan we have already started to work on is the "identification of properties". The procedure was similar to that used for the identification of classes. According to our scenarios and requirements we initially defined the needed properties in a form independent from a common metadata standard. Then we started to align these properties to properties from the RDF compliant schemas listed in chapter 3.4. Properties and classes will then be used to develop domain specific application profiles and profiles for the hubs in step six. The abstract representation of interlinked entities and the characterization of the various relationships in the model will turn into substantiality by testing the model's functionality with concrete data describing objects known to have the provenance Baron von Asch. The tests will be carried out in various annotation systems and are defined as the last step on our way to develop the ASCH model.

References

Baca, Murtha (ed.; 2008): Introduction to Metadata. Los Angeles: The Getty Research Institute.

Berners-Lee, Tim (27 July 2006): Linked Data Design Issues. 27 July 2006. W3C-Internal Document. http://www.w3.org/DesignIssues/LinkedData.html.

DiNucci, Darcy (1999): "Fragmented Future". In: Print 53,4; pp. 32, 221.

Gilliland, Anne J. (2008): "Setting the Stage". In: Baca, Murtha (ed.): Introduction to Metadata. Los Angeles: The Getty Research Institute; pp. 1-19.

Hauser-Schäublin, Brigitta; Gundolf Krüger (eds.; 2007): Siberia and Russian America: Culture and Art from the 1700s. The Asch Collection Göttingen. München, Berlin, London, New York: Prestel.

O'Reilly, Tim (09/30/2005): What is Web 2.0. O'Reilly Media. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. Retrieved 2016-08-19.

W3C (2015): Semantic Web. https://www.w3.org/standards/semanticweb/. Retrieved 2016-08-19.

Zeng, Marcia Lei; Jian Qin (2008): Metadata. London: Facet Publishing.

Zeng, Marcia Lei; Jian Qin (2008): "Schemas – Structure and Semantics". In: Metadata. London: Facet Publishing; pp. 87-130.

Wood, David; Marsha Zaidman, Luke Ruth with Michael Hausenblas (2014): Linked Data. Structured Data on the Web. Shelter Island, NY: Manning.

Aggregating Metadata from Heterogeneous Pop Culture Resources on the Web

Senan Kiryakos
University of Tsukuba, Japan
senank@gmail.com

Shigeo Sugimoto
University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp

Mitsuharu Nagamori
University of Tsukuba, Japan
nagamori@slis.tsukuba.ac.jp

Tetsuya Mihara
University of Tsukuba, Japan
mihara@slis.tsukuba.ac.jp

Abstract

Japanese pop culture resources, such as manga, anime, and video games, have recently experienced an increase in both their consumption, and appreciation for their cultural significance. Traditionally seen as solely recreational resources, the level of bibliographic description by cultural heritage institutions has not kept up with the needs of users. In seeking to remedy this, we propose the aggregation of institutional data, and rich hobbyist data sourced from the web. Focusing on manga, a form of Japanese comic, this paper discusses classification and aggregation, with the goal of improving bibliographic description through the use of fan created data. Bibliographic metadata for manga was collected from the Japanese Agency for Cultural Affairs media arts database, along with several English language manga fan websites. The data was organized into classes to enable property matching across data providers, and then tested with existing ontologies and aggregation models, namely Europeana and the Open Archives Initiative's Object Reuse and Exchange, to determine their suitability in working with these unique resources. The results show that existing ontologies may be suitable for use with pop culture materials, but that new vocabulary terms may need to be created if there is an abundance of granular data that existing ontologies fail to properly describe. In addition, the OAI-ORE aggregation method proved to be more promising than EDM when examining the aggregation of related pop culture resources. The paper discusses these issues, as well as recommendations for addressing them moving forward.

Keywords: metadata; pop culture resources; Manga; metadata aggregation model; OAI-ORE

1. Introduction

Resources that fall under the umbrella of Japanese popular culture, such manga – Japanese comics, anime – Japanese animation, video games, and others, are important pieces of cultural heritage that have historically been treated as less significant than more traditional materials by various memory institutions. As both appreciation for the cultural significance and overall consumption of these resources is increasing, the need has arisen for improved resource description and representation of these materials by institutions that collect them.

While traditional cultural heritage institutions have historically created the bare minimum in descriptive metadata for pop culture resources, special institutions and hobbyist data providers have given the materials more attention. These institutions and providers range from libraries focusing on collecting materials from a single pop culture medium, to web resources such as Wikipedia or fan websites, the latter often containing the most abundant and granular information available for a given resource. Thus, using Linked Data concepts and technologies, there is an opportunity for memory institutions to improve the state of how their pop culture resource collections are represented using already existing, hobbyist created data from the web.

This paper outlines the exploration of this data sharing opportunity. Focusing on Japanese manga, the paper examines how resources are modeled between different data providers, specifically the Japanese Agency for Cultural Affairs and various non-Japanese fan websites, and

proposes a unified model for both. The discussion then moves on to aggregation, with an examination of current metadata models, namely Europeana's, and their suitability in aggregating pop culture materials. The remainder of the paper is organized as follows. Section 2 details background information and project goals. Section 3 lists related works and past projects. Section 4 discusses how different data providers handle manga and how they can be more formally classed, as well as some discussion on pop culture ontologies. Section 5 focuses on aggregation, discussing the suitability of existing models to work with pop culture resources, as well as the benefits of aggregation. Finally, Section 6 concludes with some points of discussion and outlines future work.

2. Pop Culture Data Providers & Research Goals

The goal of sharing data between traditional cultural heritage institutions and hobbyist oriented fan websites is central to this research. The logic behind this data sharing is that different types of data providers describe materials in different ways, particularly in the realm of pop culture. In this paper, pop culture resources is used to refer to a specific subset of Japanese resources, namely anime, manga, and videogames. While research into all of these is being undertaken, this paper will generally focus on manga, a form of Japanese comics.

When discussing different data providers and how description for pop culture resources differs between them, two main provider types are relevant. The first are more traditional cultural heritage or memory institutions, such as libraries, or corporate bodies with a professional interest in recording bibliographic data for these resources. While the data recorded differs between these institutions, it is typically traditional properties one would find in a library catalogue meant to keep track of on-hand items for collections, or record data relevant to business practices. In this study, the Japanese Agency for Cultural Affairs, or Bunkacho (文化庁) represents this data provider type. Bunkacho maintains a pop culture database for several media types at https://mediaarts-db.jp. For their manga database, the data came from a corporate body responsible for the physical production of manga, and was created with consultations from libraries. Thus, the level of description granularity and specific bibliographic properties matches that of a traditional library catalogue record.

For more granular pop culture resource data, one must look to hobbyist resources, which typically take the form of a fan website. As these data providers are not bound by traditional cataloguing rules and are usually open to editing by users, the data recorded tends to be much more granular than the previous provider type. Properties such as character names and relationships, story arc summaries, genres and tags, etc., are commonly found at hobbyist sources and missing from cultural heritage institutions. Past studies have shown that fans of pop culture materials are interested in minutiae (Fee, 2013), but this is also demonstrated plainly by the fact that when able to record bibliographic data themselves on editable fan websites or on Wikipedia, this granular data is what they choose to record. As manga is the focus of this study, several fan sites were chosen based on their large manga databases: Manga Updates (www.mangaupdates.com), AnimeNewsNetwork (www.animenewsnetwork.com), and MyAnimeList (http://myanimelist.net).

In this early stage, naming specific applications of our research is difficult, but as we seek to aggregate data from both fan web pages and cultural heritage institutions, the aim is to serve users of both of these sectors. For cultural heritage institutions, improving the amount and granularity of data within their records through the accessing of fan site data is one obvious benefit for users of those institutions; librarians at several US universities have expressed a desire to include this data in their records to the authors, though have been unable to for a variety of reasons, such as staff workload. The use cases for fan sites is less obvious, as they tend to have more data than the other providers being aggregated, so their amount of information would not necessarily increase from aggregation. There are, however, interesting possibilities if one considers making aggregated information from fan sites available as Linked Data, such as aggregations being made

available via a URI and acting as the representation for a manga Work which any site can reference, for example. Section 5.1's discussion on the OAI-ORE aggregation model discusses this idea further.

All of that said, the goal of this research is to enable the data sharing between these two different data provider types through classification and aggregation in an attempt to achieve a more thorough bibliographic description landscape for pop culture resources and better serve the needs of users of relevant cultural heritage institutions and fan sites. In addition to improved resource description, we hope to provide extendable aggregation methods for connecting data across languages.

3. Related Research

He, Mihara, Nagamori, & Sugimoto used Wikipedia, through DBPedia articles for manga, as a method of identifying FRBR Works using Linked Open Data (LOD) resources (He, Mihara, Nagamori, & Sugimoto, 2013). The authors used DBPedia as a reference authority in order to identify Work level entities of manga in the Kyoto Manga Museum's catalogue. While this study focuses less on DBPedia, it is similar in that it needs to identify Work level manga entities from web resources in the absence of a traditional authority.

Southwick (2015) looked at the transformation of digital collections metadata from the University of Nevada, Las Vegas, into Linked Data. The motivations given for the project were the desire to break up the isolated digital collection metadata silos, to connect data from multiple providers, and to improve search capabilities. As the goal of this research is to similarly connect isolated data through transformation into Linked Data, the lessons learned and technologies used were helpful.

The authors previously conducted a similar study on manga metadata aggregation (Kiryakos & Sugimoto, 2015) that focused on an EDM based model, which also meant to aggregate manga metadata from different provider types. The end goals remain similar, though the previous model focused on the use of EDM and BIBFRAME, and was more focused on harvesting data and making it work within the developed model. This study takes a different "bottom-up" approach, with modeling the original data taking priority, and includes hobbyist resources, absent from the previous work. Still, some foundational efforts and lessons learned during the previous work remain useful.

While most early work focused on EDM as a method for aggregation, more recent examinations of the OAI-ORE model have been undertaken. Ferro & Silvello (2013) define a formal basis for using OAI-ORE as a way to model whole archives. Their desire to formally define complex relationships between resources is in line with the study undertaken here. The exploration of nested sets and compound digital objects is also of particular interest, as it resembles the aggregation of various resources to form FRBR-like entities, such as a complex Work entity.

4. Bibliographic Metadata for Manga

This section describes the manga metadata sourced from the previously named data providers, as well as an attempt to classify them. Both Bunkacho and the web resources have uniform properties across their respective pages, though both lack a formal data model or classification structure. In order to properly map aggregate and map data across these provider types, some classification is therefore necessary. The section ends with a discussion on pop culture specific ontologies.

4.1. Bibliographic Data for Manga

The Bunkacho manga database, located at https://mediaarts-db.jp/mg/, catalogues three main entity levels for manga. While the data can be accessed directly through the website, the authors

were given access to the database files themselves. As the database was created in consultation with librarians, the data recorded is similar to that found in a traditional library catalogue record, though lacking any formal structure or vocabulary, e.g. RDA, MARC, etc. The Comic Works pages (ex. https://mediaarts-db.jp/mg/comic_works/XXX) represent the conceptual FRBR *Work* level for a manga. These pages contain a small number of bibliographic properties based on the manga, but represent the conceptual *Work* as they contain links to related

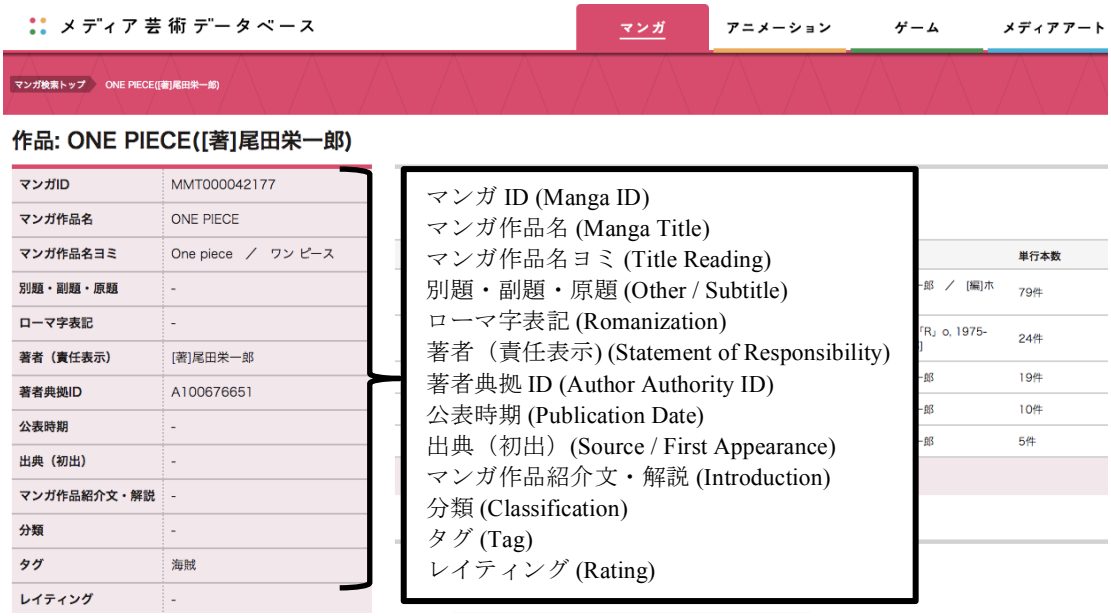


FIG 1. Screenshot of the Bunkacho Comic Works entry for the manga One Piece and added translations. Full page at https://mediaarts-db.jp/mg/comic_works/81200

entities in the database that are not manga, such as related anime entries. Therefore, while most of the data in these pages is based off of the manga *Expression* of the *Work*, they can still be seen as "home pages" representing the *Work* concept. An example of the Comic Works page is shown in Figure 1. Below this level are the Book Titles pages (ex. https://mediaarts-db.jp/mg/book_titles/XXX), which represent a combination of the FRBR *Expression* and *Manifestation* entities. If the Comic Works pages represent the conceptual *Work*, the Book Titles pages act as representations of the manga specifically, containing more manga-specific bibliographic data, as well as individual volume names and numbers. Despite most of the metadata on these pages being broad enough to apply to the *Expression* level, some properties are based on specific publication instances, thus representing the *Manifestation* levels as well. Lastly, the Books pages (ex. https://mediaarts-db.jp/mg/books/XXX) describe individual manga volumes, representing the FRBR *Manifestation* and *Item* levels. While most of the bibliographic data for the Books pages are at the *Manifestation* level, they contain some *Item* level data based on holdings information for a number of Japanese libraries. In regards to the size of the database, the amount of records vary depending on media type and FRBR entity level, but it is quite sizeable, with records for individual titles numbering over 80,000. This number includes separate editions, however, so the number of unique *Work* entities is closer to the 30,000 range.

As the Bunkacho database properties are similar to those found within traditional library catalogues, the data resembles existing generic bibliographic description models, with some changes. For example, prior to being published as volumes, manga are initially published as individual chapters in magazines at weekly or other regular intervals; a model designed to better represent manga, then, needs to incorporate these magazines and their relationships to volumes.

Manga metadata from fan sites – an example of which can be seen in Figure 2 – is quite different, with many of the bibliographic properties being unique to these data providers. This specific, granular data is the basis behind the goal of aggregating data, as it tends to be data fans of these resources are more interested in. The properties can differ between sites, but some examples that are present here and missing from Bunkacho and other traditional institutions are chapter titles, volume / plot summaries, tags and genres, character information, and spin-offs or related manga. Unlike Bunkacho, where direct database access was available, the authors had to use available APIs and HTML scraping to access data from these sources. While there do not appear to be any copyright issues, the sustainability of access, particularly through the HTML scraping method, needs to be investigated if a future project relies on constant data gathering. The number of individual *Works* handled by these sites is slightly less than the Bunkacho database, though still quite sizeable. AnimeNewsNetwork, for example, contains over 18,000 records, and MangaUpdates contains around 13,000.

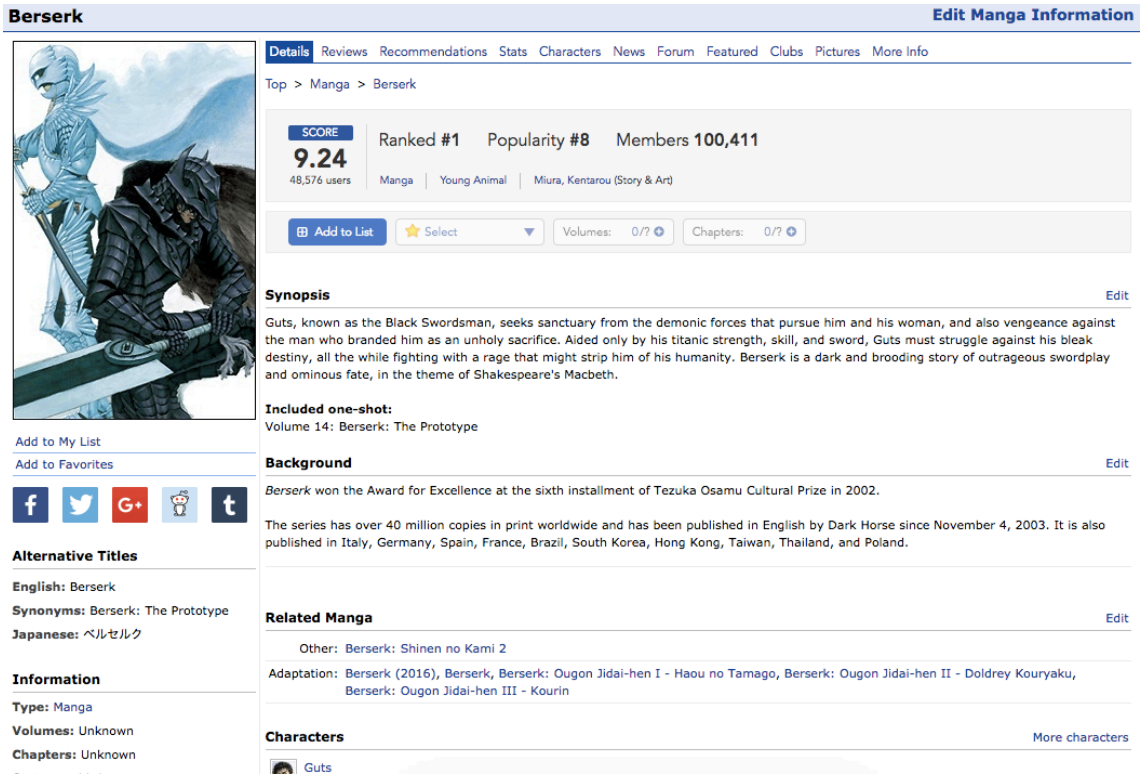


FIG 2. A sample fan site page from MyAnimeList for the manga Berserk. Full page at http://myanimelist.net/manga/2/Berserk

Importantly, the data from these different provider types is multi-lingual. While Bunkacho's database is in Japanese, the fan sites used here are English, despite containing data for Japanese resources and not English translations. As manga is a Japanese resource, Japanese data appearing in English records is common. In a previous study (Kiryakos & Sugimoto, 2015) using manga metadata from Monash University's JSC Manga Library and various US University libraries, the presence of Japanese data in the English records enabled the identification of related manga resources from Bunkacho's database when no official translation was available. Similarly in this study, the fan sites containing Japanese data enable the matching of related manga metadata across languages. While this allows for related resources to be identified and aggregated more easily at the property level, some issues arise when mapping at the class level, as will be shown in Section 4.2.

4.2. Classes for Manga Metadata

In order to more easily map the data between provider types, a self-defined class structure was created, and the properties from both Bunkacho and the fan sites were classed accordingly. This was done mainly as a preliminary process to determine how much of the data was similar across the data providers without having to do a complete mapping of all available properties across all data providers.

The majority of data from both Bunkacho and fan sites were placed into three main classes: Title, Agent, and Publication. An example of the class assignment based on the can be seen in Table 1, which shows Title class properties sourced from the various data providers. This broad class structure resembles some other bibliographic description models, such as BIBFRAME's 2.0 model (Library of Congress, 2016). In this early stage, a self-defined class structure was preferred to existing models, as it does not force the improper classification of the less traditional granular properties sourced from the fan resources, discussed further in this section.

TABLE 1: Title Class properties from Bunkacho database and manga fan sites.

Title Class	
Property	**Source**
マンガ作品名 (Manga Title)	Bunkacho
別題・副題・原題 (Other Title; Subtitle; Original Title)	Bunkacho
ローマ字表記 (Romanized Title)	Bunkacho
Title	MangaUpdates
Related Series	MangaUpdates
Associated Names	MangaUpdates
Serialized In (Magazine)	MangaUpdates
Category Recommendations	MangaUpdates
Name	AnimeNewsNetwork
Note [serialized in]	AnimeNewsNetwork
Alternative Title	AnimeNewsNetwork
Name	MyAnimeList
Serialization	MyAnimeList
Related Manga	MyAnimeList

As shown in Table 1, the mapping of properties becomes rather straightforward when organized into classes, particularly for the aforementioned main classes. Even among the Title main class, however, there are issues; for example dealing with title data that is in multiple languages can be problematic. Bunkacho data is generally in Japanese, thus the title data is in Japanese. The fan sites used in this project describe Japanese manga resources, but do so using translated English data, with Japanese data appearing as supplemental. This means that in Table 1, the Bunkacho property マンガ作品名 (Manga Title) would be mapped to the Associated Names property at MangaUpdates, rather than the Title property, which would be the English translated title. This is not a substantial issue at the moment, though it needs addressing before being able to determine how title data fits into subclasses such as Main or Alternate Title.

The main classes of Title, Agent, and Publication, along with some other classes such as an Identifier class and various subclasses, are able to contain the majority of the standard bibliographic data, particularly the data coming from Bunkacho. Classifying the more granular, manga-specific data, however is a current work in progress; properties that describe things such as character relationships or story arcs are typically absent from traditional bibliographic description models, and thus work needs to be done to create classes and subclasses that are able to logically describe this data. Petiya (2014) attempted this to an extent with comic books and

graphic novels. The resulting classes can be seen at https://comicmeta.org/cbo/. Aside from the creation of unique classes such as a "Comic" superclass, some from schema.org are used, such as schema:CreativeWork being a superclass for a ComicUniverse class. Further investigation is needed to determine whether existing classes, such as those from schema.org, are suitable for granular data for manga or other pop culture resources, or whether the creation of new classes is required.

4.3. Usability of Existing Ontologies

As Linked Data concepts are key to the sharing and description of the pop culture resources being discussed, there is the question of available ontologies that can be used to describe these unique resources.

For general bibliographic description, there are several vocabularies available for use that can sufficiently describe the majority of the properties from both Bunkacho and the fan sites. The authors' previous related study (Kiryakos & Sugimoto, 2015) experimented with the BIBFRAME vocabulary, which worked well, but has since undergone a 2.0 revision. The revision still looks to be suitable for description, but waiting for a more finalized version of BIBFRAME is advisable. Viable alternatives to this are the combination of other existing vocabularies, namely Schema.org and Dublin Core – National Diet Library (DCNDL), the latter of which is particularly useful for Japanese resources, as it already contains properties to describe data such as Japanese readings or transliterations.

For pop culture specific bibliographic properties, other issues arise. As one main goal of this research is to access the rich hobbyist data that is absent from cultural heritage institutions, one must be able to properly describe this data, particularly in a Linked Data context – in other words, using RDF based vocabularies. The vocabularies to describe this data, however, generally do not exist. There has been some work in creating some resource specific ontologies for pop culture resources (Petiya, 2014), but how extendable these are to other resource types needs to be examined. Petiya's ontology, for example, is rather thorough if one wants to describe American comic books and aspects of collecting them, but may be unable to adequately describe nuances of mediums in the Japanese pop culture sphere, such as diverse manga publication hierarchies or relationships to common spinoff media types such as videogames or film. Ultimately, the amount of granular data available will determine the necessity of these unique vocabularies; if there exists enough data that is unique to manga, anime, etc., that institutions would like to access, at least some new properties will have to be created to accommodate this. The majority of the data being described, however, is fairly standard with other bibliographic materials, so the opportunity to reuse parts of existing ontologies is present and will no doubt be performed.

5. Aggregating Pop Culture Data

This section discusses the suitability of existing aggregation models for pop culture data, and demonstrates some of the benefits of aggregation based on the manga example discussed throughout the paper.

5.1. Aggregating Using Existing Models

In connecting hobbyist and institutional data, and underlying aggregation model is required. Perhaps the most prevalent existing linked data aggregation model is the Europeana Data Model (EDM). The intended use of EDM is for aggregating cultural heritage data sourced from various European institutions for display on the Europeana web portal. Typically these objects are singular, unique items that exist in galleries or museums, and so the use of EDM with non-unique items such published manga is not as straightforward. The use of EDM with manga was previously examined (Kiryakos & Sugimoto, 2015) to determine how compatible pop culture materials were with the model. While EDM was suitable for some tasks, when it came to representing the various FRBR entities that are found with manga data, the authors found EDM to

be less than ideal; similar issues were noted in an EDM for Libraries document (Angjeli et al., 2012).

The initial issue was with the use of the mandatory edm:ProvidedCHO class. In EDM, the CHO (Cultural Heritage Object) represents the, typically unique, object being described, such as a painting or sculpture. While determining what the CHO represents when discussing a unique object such as a sculpture is usually clear, it is less so when dealing with objects such as literary materials with numerous copies, editions, publications, etc. Angjeli et al (2012) found that there was some confusion on whether to apply edm:ProvidedCHO to the specific *Item* of a textual resource, or the "edition level" representing FRBR's *Work*, *Expression*, and *Manifestation*. Consultation with Europeana revealed that a ProvidedCHO could represent both the *Item* and edition levels, with metadata establishing a relationship between the two, e.g. the edm:ProvidedCHO for an item is connected to the edm:ProvidedCHO of the edition through an edm:realises property. For a resource such as manga, however, the representation of multiple FRBR entities seems less than ideal due to the amount of relationships a single resource can have to related resources. For example, a single volume (*Item*) of manga can be connected to the manga series to which it belongs. This can then be connected to translation of the work, and both of these could in turn be connected to the manga *Expression* to which they belong. The *Work* level entity as well can connect the manga to an anime adaptation, something common among Japanese pop culture resources. Using EDM means that all of these different entities are represented as the same edm:ProvidedCHO class, and are connected through limited relationship properties such as edm:realises. As the authors wish to adequately model and describe relationships between pop-culture resources at each level of the FRBR entity hierarchy, a model that uses more than a single property to represent multiple FRBR entities is preferable.

Another promising method separate from EDM being investigated is the use of the OAI-ORE aggregation model (Lagoze et al., 2008). This model allows for the creation of a RDF-based Resource Map is created that describes an aggregation of existing web resources. For example, a series of web documents that all describe a single manga volume can conceptually be considered an aggregation, and a Resource Map can be created that asserts some amount of information about that aggregation. These aggregations can themselves be aggregated, possibly allowing for the portrayal of FRBR-like hierarchies within the model (i.e. a group of resources are aggregation for the *Expression* level, with multiple *Expression* level aggregations forming a *Work* level aggregation). As the Resource Map is given a URI, one possible outcome of this model is to create a Resource Map containing metadata based on aggregated resources, and use the URI as the web representation of whatever that aggregation may be describing. The feasibility of this will be determined in the near future, but the idea is a promising one.

While the best option moving forward still has yet to be determined, the OAI-ORE method may be the most suitable, at least when compared to previous efforts using EDM. It more readily utilizes existing web resources, and allows the application of bibliographic metadata to aggregations representing multiple FRBR entity types. The creation of Resource Maps representing different FRBR entity levels for pop culture resources may also enable interest web applications. Issues such as how easily this data can be automatically created, whether or not this would require much data harvesting, and what types of relationships between resources can be asserted within the aggregation model first need to be investigated before a more certain future path can be established.

5.2. Benefits of Metadata Aggregation for Pop Culture Resources

As stated previously, the goal of aggregating pop culture data, specifically using hobbyist resources such as fan sites, is to improve the granularity of data that is available for these resources. Like the Europeana portal, this means providing data for the same objects that comes form multiple perspectives, and multiple languages. The use of resources such as fan sites also enables describing the minutiae of pop culture resources, which are typically absent from traditional cultural heritage institutions. It also enables the filling in of gaps for bibliographic data

that institutions may have attempted to record, but remain blank for a variety of reasons. Figure 1 illustrates this, with several of the left sidebar properties containing no data. Fan sites can be utilized in a pseudo-crowdsourcing fashion to remedy this, helping to improve the amount of useful data in institutional records with data created by hobbyists.

Figure 3 shows an example of Bunkacho properties (middle) that are commonly unfilled, and grouped properties from previously mentioned manga fan sites that contain data that could be used as a supplemental to fill in the missing data. Thus, even if a traditional cultural heritage institution is unable to utilize some of the more granular data, there remains the opportunity to essentially have their missing data "crowd sourced" thanks to existing hobbyist resources.

While the idea remains to be more thoroughly examined, there is also the opportunity for fan sites to use an aggregated database centered around cultural heritage institutions as a type of pseudo-authority for pop culture resources. Existing library authorities are suitable for titles and creator names, but most other facets of these resources are inadequate; one simply needs to look at how existing Library of Congress Subject Headings terms are used to describe manga to understand this. Creating authorities based on institutions such as Bunkacho and the Kyoto International Manga Museum, and supplementing them with hobbyist data, would be beneficial for all parties and their users. Also, as mentioned in Section 5.1, an OAI-ORE method of aggregation may enable the creation of a Resource Map with a URI that can act as a web representation of a pop culture resource instance, allowing hobbyist sites to include their information in the Linked Data cloud once aggregated. These unique Resource Map URIs and the bibliographic metadata they contain could potentially act as the content of the pop culture "pseudo-authority".

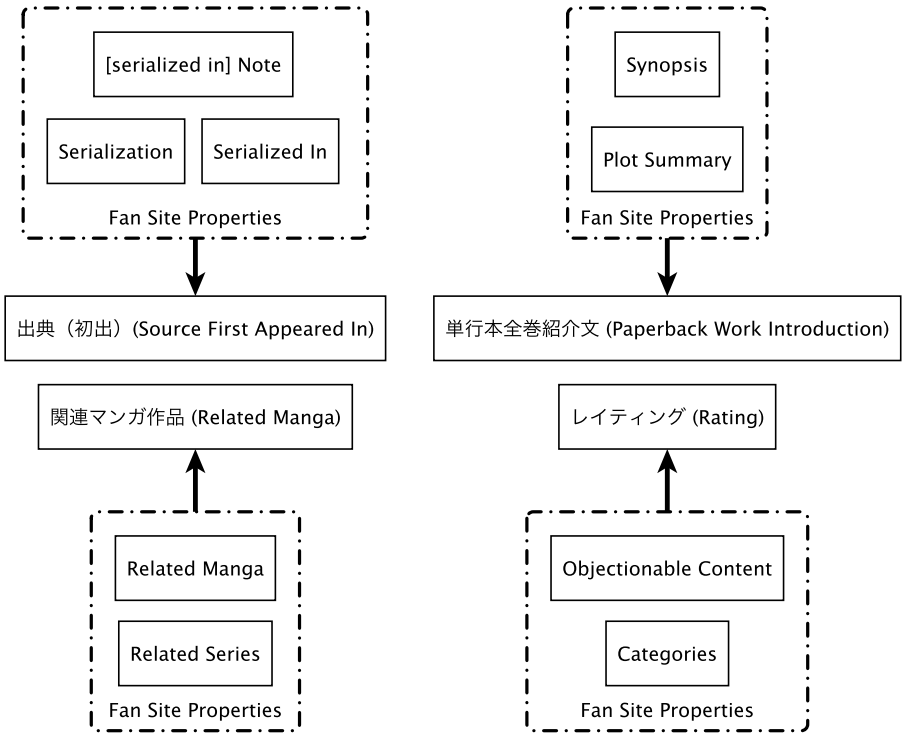


FIG. 3. Bunkacho properties that are typically lacking any data (middle) and grouped fan site properties that contain suitable data for those properties.

6. Conclusion and Future Work

This paper has outlined the preliminary work that has gone into aggregating pop culture metadata from traditional cultural heritages and hobbyist resources. Through the collection,

classification, and mapping of manga metadata, as well as the examination of how this data fits into existing aggregation models, the foundation for future work has been established. The inclusion of hobbyist data also builds on prior work, and makes available data that fans of these resources are genuinely interested in accessing.

There remains much work to be done, however, to realize the goal of improved pop culture resource description through metadata aggregation. As much of the hobbyist data being accessed is unaccounted for in traditional bibliographic description models, a formal classification scheme that is able to model this data should be created. Similarly, the ability for existing vocabularies to describe pop culture specific data must be investigated more thoroughly, as representing this data in an aggregation model is dependent on useable properties. Mentioned in Section 5.2, the ability for EDM to accurately represent the relationships between pop culture resources is questionable, particularly when one wants to aggregate data from multiple media formats, so a solution to this too needs to be investigated, be it through the use of alternative models or the development of new aggregation properties. The OAI-ORE aggregation method has just begun to be investigated for use with these materials, but it appears to be a promising alternative to our previous work, particularly in regards to possible use cases. This seems to be a practical basis for the creation of a web authority for pop culture resources, which would no doubt improve the information sharing landscape for these resources. While manga has been the focus of this and past research, the authors would like to experiment with the inclusion other mediums, such as anime and video games, as these resources typically have multiple explicit relationships between manga and each other. If future projects are able to address these issues, then the aggregation of hobbyist and institutional data should provide for an improved bibliographic description landscape for a wide variety of related pop culture resources.

Acknowledgements

Thank you to everyone in the MDLab for assistance and suggestions throughout our research. In addition, thanks to the Japanese Agency for Cultural Affairs for access and use of their data. This study is supported in part by the JSPS Kaken Grant-in-Aid for Scientific Research (A) #16H01754.

References

Angjeli, Anila., et al. (2012). D5.1 Report on the alignment of library metadata with the Europeana Data Model (EDM).

Fee, William. T. B. (2013). Where Is the Justice… League?: Graphic Novel Cataloging and Classification. Serials Review, 39(1), 37–46. doi:10.1080/00987913.2013.10765484

Ferro, Nicola. & Gianmaria Silvello. (2013). Modeling Archives by Means of OAI-ORE. Digital Libraries and Archives, 216–227. doi:10.1007/978-3-642-35834-0_22

He, Wenling, Tetsuya Mihara, Mitsuharu Nagamori & Shigeo Sugimoto. (2013). Identification of works of manga using LOD resources. Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '13. doi:10.1145/2467696.2467731

Kiryakos, Senan., & Shigeo Sugimoto. (2015). A Linked Data Model to Aggregate Serialized Manga from Multiple Data Providers. Lecture Notes in Computer Science, 120–131. doi:10.1007/978-3-319-27974-9_12

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., & Warner, S. (2008). ORE Specification-Abstract Data Model. http://www.openarchives.org/ore/datamodel

Library of Congress: Overview of the BIBFRAME 2.0 Model. (2016). Retrieved from https://www.loc.gov/bibframe/docs/bibframe2-model.html

Petiya, Sean. (2014). *Building a Semantic Web of Comics: Publishing Linked Data in HTML/RDFa Using a Comic Book Ontology and Metadata Application Profiles.* (Electronic Thesis). Retrieved from https://etd.ohiolink.edu/

Southwick, Silvia. B. (2015). A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies. Journal of Library Metadata, 15(1), 1–35. doi:10.1080/19386389.2015.1007009

Automatic Creation of Mappings between Classification Systems for Bibliographic Data

Magnus Pfeffer
Stuttgart Media University, Germany
pfeffer@hdm-stuttgart.de

Abstract

In this paper, the implementation of an approach to automatically create mappings between classification systems is presented and results from a preliminary analysis are discussed. The approach is based in the idea of instance-based ontology matching and consists of three steps: First, bibliographic data from diverse sources that contain items classified by the required classification systems is aggregated in a single database. Next, an efficient clustering algorithm is used to group individual issues and editions of the same work. It works by matching names of authors and corporate bodies as well as title, subtitle and uniform title. Finally, the clusters containing information from both required systems are added up to create a co-occurrence table. This information is then used to generate candidates for a mapping between the individual classes of the two classification systems.

In an experiment, the implementation is utilized to generate mappings between two classification systems that are in use in Germany. The mappings are evaluated using existing partial mappings that have been manually created by domain experts as a gold standard for comparison. While the automatic mappings might be less accurate and exhaustive than manually created ones they are sufficient for retrieval and visualization purposes and could be further improved by refining the statistical analysis or including more datasets.

Keywords: library catalog, classification systems, instance-based ontology mapping

1. Introduction and Motivation

Classification systems are an important means to provide topic-based access to library collections. Depending on the collections at hand and the primary use cases, these classification systems can differ significantly in structure and organization. For example, systems used to arrange large collections on shelves need to be sophisticated and highly structured in order to keep the number of members of each class manageable. On the other hand, applications like topic-based facetted browsing in resource discovery systems or graphical representations of the contents of collections benefit from a simpler structure with fewer branches and depth to assure a clearly arranged presentation to the user. With the proliferation of more powerful search solutions in libraries, there is a renewed interest in using different classification systems for search or browsing.

As annotating a library collection using multiple classification systems would be prohibitive, using mappings to derive new annotations from existing data is a possible solution. The creation of such mappings can be an arduous process if done manually, but is still undertaken for applications in information retrieval systems or to assist library collection reorganization. Part of the ongoing projects of the Austrian National Library, as presented in Plößnig (2012) and Plößnig (2014) is the enrichment of the catalog data with annotations using multiple classification systems. For this purpose several partial mappings from the *Regensburger Verbundklassifikation* (RVK, engl.: Regensburg union classification system) to the the *Basisklassifikation* (BK, engl.: basic classification system) have been created manually and are already used to enrich catalog entries.

In this paper, we propose an automated approach to automatically create mappings between classification systems used in libraries. It is based on the idea of instance-based ontology matching, which works on the annotated instances instead of comparing the labels of classes or the structure of the systems. The general applicability of this matching method to data from library catalogs has been shown in multiple projects in the past (Isaac et.al., 2007, Schopman, 2009 and Schopman et. al., 2012) and in own prior work, preliminary data generated from the implementation was used as input for the manual mapping project at the Austrian National Library with positive results (Aigner, 2005). The approach is tailored to library catalog data with its specific properties and its implementation prepared to scale up to very large datasets with more than 100 million entries.

To evaluate the results of the mappings process and to create a baseline for further experiments, a large dataset of catalog data containing entries annotated with RVK and BK classes has been collected and a full mapping was produced using the proposed approach. A relevant subsection of the existing manual mapping results from the projects of the Austrian national Library was selected to be used as a gold standard to evaluate the automated mapping.

The paper is structured as follows: First is a short review on the different methods of ontology matching and the related work on instance-based ontology matching in the library domain as well as prior work of the author that has influenced the development process. Next the implementation specifics of the approach, the design decisions and their inherent tradeoffs are discussed. The second half of the paper focusses on the evaluation: the used datasets and classification systems are introduced in detail and the resulting automated mapping is compared to the gold standard by calculating precision and recall for a range of parameters. The paper closes with a look at further possible enhancements to the approach itself and the current software implementation.

2. Preliminaries and Related Work

Ontology mapping is a vast and very active field of research with many applications in knowledge organization and knowledge representation, especially for the Semantic Web. While the ontologies discussed in this field are often very rich structures expressed in OWL or similar high level languages, there is also an interest in less rich ontologies that can be expressed in SKOS or data formats traditionally used in library information systems. The Ontology Alignment Evaluation Initiative[1] regularly invites participants to compare and benchmark their latest algorithms and includes a library track specifically for this kind of data since 2012 (Aguirre et. al., 2012).

Euzenat and Shvaiko (2007, p. 341) lists four automatic ontology matching methods: terminological, structure-based, semantic-based and instance-based. Terminological methods work on the lexical data contained in concept labels or descriptions and utilize it to find matches by string comparison. Structure-based methods use the relations between concepts to deduce possible matches. Semantic-based methods use generic or domain-specific rules or other background information outside the ontologies being matched. Instance-based then rely on the set of instances that are associated with a given concept. Depending on what type of instances are available, different methods can be applied: If instances exist that are annotated using two ontologies, one can directly analyze the co-occurrence of concepts; the idea being that two concepts are closer related, the more significant the overlap of common instances of two concepts is.

If no such dually annotated instances exist, it is possible to extend the concepts themselves using the contents of the annotated instances and compare these extended concepts. Alternatively one can try to match the instances themselves and create clusters of instances, which are then again the basis for a co-occurrence analysis.

[1] http://oaei.ontologymatching.org/

Instance-based matching has advantageous properties: it is less affected by ambiguities like homonyms or synonyms that are inherent in limited lexical data like labels or short description. Also as the sets of instances are the result of practical application of the ontology on documents, they are a very precise representation of the concepts true meaning. Finally, the method can cope with small annotation errors or variances that are inherent to a manual annotation process that is done by several individuals. On the other hand, it is often difficult to find sufficient instances, i.e. annotated objects or documents.

Instance-based ontology matching has been successfully implemented and used with data from libraries in the past: Isaac et al. (2007) created a mapping between a classification system and a thesaurus based on data from the Dutch National Library and evaluated the result by comparing to an existing manual mapping. In Schopman (2009) this work was extended to include multilingual data from the European Library and the algorithms were further refined. Both reports showed very encouraging and positive results. Finally, in a paper by the same authors, the algorithm and application is further generalized and rigorously evaluated it using large multilingual data sets (Schopman et. al., 2012).

In the library domain, finding a large number of instances is less problematic, as most libraries seek to enable a topic-based search or access by using a classification system or thesaurus to annotate the catalog entries. Nonetheless, it is often not the case that catalog entries are uniformly and consistently annotated: the use of ontologies can change over time or resources may be insufficient to keep up with manual annotations. In Germany, there is an additional complication: due to historic developments, there are several large library unions, each with their own central cataloging database alongside the National Library with its own catalog. Data sharing between these entities has been limited and resulted in very heterogeneous data sets, especially in regards to annotations using classification systems. The author has applied different clustering methods on data sets from German library unions in order to enrich entries with annotations from other library unions and evaluated the results using existing manual annotations as gold standards (Pfeffer, 2009). One important result was that generic clustering methods like k-nearest neighbor based on string similarity tend to create inconsistent clusters, resulting in a low precision for the enrichments, while clustering based on exact matches of title and subtitle and author/corporate bodies resulted in very consistent clusters and very high precision for the enrichments, which was considered to be on par to most manual annotation by indexing experts (Pfeffer, 2013).

Data from these enrichment projects was used to evaluate the usefulness of co-occurrence analysis for the creation of mappings in theses by library science students: In Probstmeyer (2009), a mapping between the Schlagwortnormdatei (SWD, a subject headings authority file used by most German academic libraries) and the RVK was evaluated. Co-occurrence was calculated using the individual catalog entries, and the evaluation showed that the significance of the co-occurrence was not strongly correlated with the relation of the concepts. One reason was that in the catalog data, works with many different editions tended to have the same co-occurring annotations and overshadowed the co-occurances from works which only exist in a single edition. In Aigner (2015), the process of creating a manual mapping between the RVK and BK for the domain of geography is discussed. Here, co-occurrence was calculated using the consistent clusters and the resulting matches were used as one source of possible mappings (besides mostly manual lexical and structural analysis). The analysis showed that the significance of co-occurrence was correlated with the relation of the concepts and after choosing a suitable threshold almost all remaining mappings were deemed highly useful.

3. Data Sets and Implementation

The experiments presented in this paper are a direct result of the lessons learned in preparing the co-occurrence data that was used successfully in Aigner (2015). The implementation used was not running stable, used a lot of computing resources and did not scale well for larger datasets. Beside the performance issues, a new implementation should also be more flexible in

regards to the data used as basis as well as the ontologies to be matched. To assess the properties of the new implementation, the full automatic process was run using several very large datasets mapping the classification systems that have been the focus at the Austrian National Library. This course of action ensured that enough information is available to evaluate the resulting mappings.

In this section, first the classification systems and the data sources used in the experiment are introduced. Next the clustering process and its implementation are presented and explained using a simplified example.

3.1 Classification systems

The RVK was developed in the 1960s as a local classification system for the library of the newly founded Regensburg University. Unlike most existing German university libraries, the collections in Regensburg were planned to be mostly openly accessible by users and this influenced the structure and design of the classification system. It is a monohierarchical universal classification system modelled on the Library of Congress classification (LCC) and consists of 33 domain-specific sections that mirror the structure of German university faculties. Granularity and hierarchies in these domain-specific sections vary to a certain extent, as well as the principles used to create further subdivisions. The RVK consists of about 80.000 classes in total. (Lorenz, 2008)

The RVK has seen continued adoption by other academic libraries and is now the most used universal library classification system in the German-speaking region, being in use at more than 140 libraries. .

All class notations have a common composition: Two uppercase roman letters are followed by a three to six digit number. For example, the notation "QF 100" is from the section "Q: Economics", subclass "QF: History of Economics" and represents "QF 100: History of Economics until 500 A.D.". See figure 1 for an excerpt of the class tree view[2].

Figure 1: Excerpt from the tree view of the Regensburg union classification system

The BK was originally developed in the Netherlands by the PICA library foundation under the name "Nederlandse basisclassificatie", based on existing domain-specific classification systems used to index bibliographies. It was translated into German in 1992 and adapted by many libraries in the North German region. The BK is a monohierarchical universal classification system

[2] An online version of the full system is available at https://rvk.uni-regensburg.de/regensburger-verbundklassifikation-online

consisting of about 2100 classes that are divided into 48 main divisions. The divisions are modelled after traditional domain structures in the sciences as well as certain interdisciplinary aspects. The classes within each main division are arranged mostly by topic, less often by region or historic timespan. BK was developed as a secondary annotation system that was to be used together with thesaurus-based indexing to provide multiple ways of topical access to collections. (Schulz, 1991)

Class notations are composed of a two-digit number, a dot as a separator and another two-digit number. The first number denotes the main division, the second one the class within that division. For example the notation "15.09: History of Economics" is part of "15: History". See figure 2 for an excerpt of the class view[3].

15.08 Sozialgeschichte
Erl.: Zu verwenden für Sozialgeschichte allgemein ohne zeitliche, räumliche oder sachliche Einschränkung (z.B. Sozialgeschichte des Bürgertum). Darüber hinaus zu verwenden als Zweitnotation zu 15.25-15.96, sofern Darstellungen zu einzelnen Epochen oder Kulturräumen dezidiert sozialgeschichtliche Themen behandeln

15.09 Wirtschaftsgeschichte
Erl.: Zu verwenden für Wirtschaftsgeschichte allgemein ohne zeitliche, räumliche oder sachliche Einschränkung. Darüber hinaus zu verwenden als Zweitnotation zu 15.25-15.96, sofern Darstellungen zu einzelnen Epochen oder Kulturräumen dezidiert wirtschaftsgeschichtliche Themen behandeln

15.10 Historische Hilfswissenschaften
Hier: Historische Geographie <Geschichte>
Verw.: Archivkunde -> 06.90 (Archive, Archivkunde)
Geschichtsatlanten ohne räumliche und sachliche Einschränkung -> 15.20 (Allgemeine Weltgeschichte)
Handschriftenkunde -> 06.10-06.19 (Handschriftenkunde

Figure 2: Excerpt of the class view of the basic classification system

3.2 Data sources

Catalog title data from most German library unions is available as open data in the MARC21 format. For the project, the following catalogs were chosen:

- Gemeinsamer Bibliotheksverbund (GBV, engl.: Common Library Network). Spanning several states in northern Germany, it is the largest library union. Its catalog also includes the collection of the Berlin state library.
- Südwestverbund (SWB, engl.: Southwest German Library Union). Its member libraries are located in the states of Saarland, Baden-Württemberg and Saxony.
- Bibliotheksverbund Bayern (BVB, engl.: Bavarian Library Union). The union catalog contains the collections of libraries from the states of Bavaria, Berlin and Brandenburg.

Table 1 contains some statistics on contents and annotations of the three datasets. Non-monographic entries (like musical notes, dvds, maps, etc.) were filtered using information from the MARC21 field "Leader" and field 007. Annotations were taken from the main title data MARC21 field 084 (subfield 2 values "rvk", "bcl" or "bkl" respectively).

Table 1: Contents of the initial datasets

	All Entries	Monographic entries	Monographic with RVK	Monographic with BK
GBV	32,027,977	24,267,492	0	3,976,154
SWB	18,789,185	16,447,890	4,383,273	0
BVB	26,680,083	23,658,674	7,215,483	0

[3] An online version of the full system is available at https://www.gbv.de/bibliotheken/verbundbibliotheken/02Verbund/01Erschliessung/02Richtlinien/05Basisklassifikation/index

The catalog of the Austrian National Library contains both RVK and BK annotations. As its entries have already been enriched extensively using the results from the manual mapping projects, it was considered to be unsuitable as a data source for this experiment.

3.3 Clustering process

The clustering process is implemented using the Perl scripting language. All data is stored in a NoSQL document database back end using only the very basic features of key-value storage and access. In the implementation for the evaluation experiment, MongoDB on a 16-core server with 16 GB of RAM is used to allow fast access even for large datasets.

In the first step, the original MARC21 data is transformed into a very simple JSON-like data format containing only the most important properties: id, title, subtitle, uniform title, author, corporate entity, publisher, year of publication and the annotations of RVK, BK and the dewey classification system (DDC) as well as index terms from the Gemeinsame Normdatei (GND, engl.: common authority file) used by most libraries in the German-speaking region. Properties that can contain more than one entry, like author or the annotations are stored as lists, all other properties as string literals. The original database ids are used as the access key ids for this *data* table.

In the second step, strings are generated for each entry of the *data* table by creating combinations of all author or corporate entity list entries with the title+subtitle and uniform title. These generated strings are used as access key ids for the *key* and *keyequiv* tables. In the *key* table the corresponding ids from the *data* table are stored as a list. In the *keyequiv* table, the other strings that were generated from the same data are stored in a set. Table 2 shows the resulting entries for a simplified example. Although the entries with 1 and 3 do not share an author, they should become part of the same cluster because they both share author and title with id 2.

To generate the clusters, in the third step the *key* table is traversed: The current id is stored in a set named "done" and all equivalent strings are retrieved from *keyequiv* and stored in a set "todo". As long a "todo" still contains entries, the first entry gets moved from "todo" to "done" and the equivalent strings for it are retrieved from *keyequiv* and stored in "todo" unless they are already contained in "done". Finally each entry of "done" is marked in *key* and the corresponding data ids from *key* are retrieved and stored in a temporary set, which is then saved as a new entry in the *cluster* table. The traversal continues with the next non-marked key in *keys*.

Table 2: Example tables illustrating the MongoDB implementation

data table	key table	keyequiv table
id: 1 author: [A, B] title: beer year: 1990	id: A|beer ref: [1]	id: A|beer eq: [B|beer]
id: 2 author: [B, C] title: beer year: 1995	Id: B|beer Ref: [1, 2]	id: B|beer eq: [A|beer, C|beer]
id: 3 author: [C] Title: beer Year: 1999	Id: C|beer Ref: [2,3]	id: C|beer eq: [B|beer]

The combination of fields to create keys in step 2 can be changed, thus influencing the resulting clusters. For this experiment, only authors, corporate bodies, uniform title and main title have been used. By ignoring the subtitles, the clustering is more aggressive and creates larger clusters, which can in theory lead to more inconsistent clusters. Earlier experiments had shown

that this happens rarely in practice, as the combination of a short title consisting of a common word or phrase and two authors with the same name is highly unlikely.

Applying the clustering process to the data sources results in 21,653,606 clusters, of which 904,876 reference catalog entries that contain BK and RVK annotations. The co-occurrence data was then generated and for each pair of BK and RVK classes that occurred at least in one cluster, the final table containing the RVK class notation, the number of dually annotated clusters that were annotated with this RVK notation, the BK class notation that co-occurred, and the number of dually annotated clusters that were annotated with the exact pair. Co-occurrence data for 1,155,552 such pairs was found.

The whole process ran very stable and reliably, using only a small part of the server resources. The whole process, from importing the data sets to finished co-occurrence table took less than 3 days.

4. Evaluation

To assess the quality of the co-occurrence data and to determine possible thresholds to filter the data, an existing manual mapping from RVK to BK for the domain of economics was chosen for comparison. The mapping was provided by the Austrian National Library and was done by Andreas Waldhör, who had done a mapping for the domain of law as a Master's thesis (Waldhör, 2012). It contains 963 individual mappings from the "Q: Economics" division of the RVK to the BK; mapping each RVK class to exactly one BK class. The corresponding selection from the co-occurrence data contains 44710 pairs, with the strongest co-occurrence being 3195 clusters sharing a specific pair.

Of the 963 manual pairs, 808 were also found in the co-occurrence list, resulting in a maximum recall of 0.839 with a precision of 0.018. Of the missing 155 pairs, only 14 contained RVK classes that were completely missing in the co-occurrence data, while the RVK classes of the other 141 pairs appeared in co-occurrence, but with different BK classes.

Two parameters were selected for filtering the raw co-occurrence data: first, the ratio of the number of clusters with a given pair to the number of pairs containing the same RVK class and second, the absolute number of clusters with a given pair. The first is a Jaccard-like measure with a maximum of 1, when all clusters that contain the RVK class from a given pair also contain the BK class. The ratio is smaller, the more clusters with the same RVK class but different BK classes exist. It was preferred over the classic Jaccard measure, i.e. the ratio of the number of clusters with a given pair to the number of pairs containing the RVK class *or* BK class of the pair, because of the imbalanced size and structure of the two classification systems being mapped: As RVK contains far more classes, any BK class is expected to be correctly mapped to a high number of RVK classes. Including the number of pairs with the BK class as well would have led to significantly higher numbers, which would in turn result in very small ratios that are harder to compare. With the goal of a mapping from RVK to BK (and not vice versa) in mind, the chosen ratio was considered to be far superior.

The second parameter can be used to filter pairs that only occur in few clusters. Tables 3 and 4 contain the precision and recall results for a range of values for both parameters. The results are decent, but not overly impressive. It is interesting to see that increasing the required number of clusters results in a significant increase in precision while the recall is not affected very much. The ratio on the other hand affects both precision and recall, with a quickly decreasing gain on precision for ratios of 0.6 and more.

Table 3: Precision results. Values >0.5 are highlighted

	ratio ≥ 0	**ratio ≥ 0.1**	**ratio ≥ 0.2**	**ratio ≥ 0.3**	**ratio ≥ 0.4**	**ratio ≥ 0.5**	**ratio ≥ 0.6**	**ratio ≥ 0.7**	**ratio ≥ 0.8**
num ≥ 0	0.0181	0.1639	0.2177	0.2410	0.2383	0.2015	0.1759	0.1100	0.0553
num ≥ 2	0.0183	0.1979	0.2979	0.3769	0.4436	0.4236	0.6218	0.6319	0.5333
num ≥ 4	0.0179	0.2129	0.3499	0.4989	0.5918	0.6269	0.7067	0.7288	0.7143
num ≥ 6	0.0173	0.2222	0.3954	0.5177	0.6353	0.6724	0.7525	0.7714	0.7561
num ≥ 8	0.0171	0.2308	0.4053	0.5280	0.6529	0.6951	0.7814	**0.8125**	**0.8056**
num ≥ 10	0.0167	0.2386	0.4089	0.4206	0.6603	0.7066	**0.7877**	**0.8261**	**0.7941**

Table 4: Recall results. Top 5 values are highlighted

	ratio ≥ 0	**ratio ≥ 0.1**	**ratio ≥ 0.2**	**ratio ≥ 0.3**	**ratio ≥ 0.4**	**ratio ≥ 0.5**	**ratio ≥ 0.6**	**ratio ≥ 0.7**	**ratio ≥ 0.8**
num ≥ 0	**0.8390**	0.6947	0.5940	0.4922	0.3801	0.2835	0.1817	0.0987	0.0457
num ≥ 2	**0.8349**	0.6906	0.5898	0.4881	0.3759	0.2793	0.1776	0.0945	0.0415
num ≥ 4	**0.8089**	0.6646	0.5639	0.4621	0.3583	0.2617	0.1651	0.0893	0.0363
num ≥ 6	**0.7809**	0.6366	0.5358	0.4403	0.3364	0.2451	0.1547	0.0841	0.0322
num ≥ 8	**0.7653**	0.6210	0.5265	0.4309	0.3281	0.2368	0.1485	0.0810	0.0301
num ≥ 10	0.7487	0.6044	0.5130	0.5315	0.3229	0.2326	0.1464	0.0789	0.0280

In order to get threshold values that balance precision and recall, f-measures were calculated. Table 5 contains the results for the f-measure, with double weighted precision. The higher weight for precision was chosen with the intended use cases in mind: using the mapping for enrichment in catalogs or as a basis for creating manual mappings would be significantly negatively affected by low precision results, and less by low recall results.

Table 5: f-measure, with double weighted precision. Top 5 values are highlighted

	ratio ≥ 0	**ratio ≥ 0.1**	**ratio ≥ 0.2**	**ratio ≥ 0.3**	**ratio ≥ 0.4**	**ratio ≥ 0.5**	**ratio ≥ 0.6**	**ratio ≥ 0.7**	**ratio ≥ 0.8**
num ≥ 0	0.0270	0.2322	0.2991	0.3221	0.3090	0.2566	0.2124	0.1290	0.0637
num ≥ 2	0.0273	0.2770	0.3968	0.4739	0.5138	0.4607	0.4974	0.3548	0.1899
num ≥ 4	0.0267	0.2957	0.4544	0.5893	**0.6283**	0.5881	0.5121	0.3596	0.1810
num ≥ 6	0.0258	0.3066	0.5007	0.6001	**0.6473**	0.5983	0.5093	0.3514	0.1651
num ≥ 8	0.0255	0.3168	0.5098	**0.6063**	**0.6540**	0.6014	0.5062	0.3474	0.1571
num ≥ 10	0.0249	0.3258	0.5114	0.5267	**0.6554**	0.6024	0.5038	0.3425	0.1472

One question remained: What kind of mappings have a highly significant co-occurrence yet are not part of the manual mappings? In an additional analysis step, the co-occurrence data was filtered by rather high thresholds of a ratio larger or equal than 0.6 and a number of clusters larger or equal than 6 and again compared to the manual gold standard. The 49 mapping pairs that were not contained in the manual list were individually assessed using the class descriptions and classification system structure.

Of the 49 mapping pairs, 31 were considered to be correct, 12 partially correct, 1 false and 5 contained RVK classes that are no longer in active use. In this sample, most of the "correct" mappings were for RVK classes for the history of economics of specific countries, which were mapped to the BK classes representing the history of those countries. In the manual mapping, there was only a descriptive note for these classes, but not an exhaustive mapping for each country. This is a clear shortcoming of the manual gold standard, that was not obvious in the beginning of the analysis. Another example is the RVK class "QP 624: product and product range selection" (a subclass of "QP 620 - QP 624: demand management instruments" being mapped to BK class "85.40: marketing" instead of the manual choice of "85.15 research and development (economics)". The manual choice was probably caused by a misunderstanding of the German labels "Produktgestaltung" vs. "Produktentwicklung" (product design and product development). The structural analysis indicates that this topic belongs to the field of marketing, so the automatic mapping can be considered the superior match.

This preliminary first analysis shows that the approach has a high potential to further improve and augment the existing manual mappings as well as create automatic mappings that can be used to improve the retrieval in resource discovery systems or be used as a first draft for manual mapping projects.

5. Discussion and Future Work

The current analysis is limited and will need to be significantly extended to more closely follow the work of the other research groups, especially in regards to the effect of different statistical measures used to select the co-occurrences. Nonetheless, several important goals for the current project have been accomplished: the implementation is fast, very robust and can handle large datasets with ease. The evaluation of the approach against the manual mapping gave decent results for precision and recall, and the in-depth analysis showed that many of the automatic mappings "false positive" pairs were actually correct and can be used to significantly improve the existing mapping.

On a more practical side, work is ongoing to document the data management pipeline and switch it over to a more maintainable and user-friendly solution based on the Knime.org framework as well as implementing the statistical analysis directly on top of the data in the back-end database. Also, the manual mappings are currently only provided on request by the Austrian National Library and are contained in Excel files with a varying layout and degree of mapping granularity. The author intends to convert them into a single, well documented format and work together with the original authors to publish them in an open data repository. The same format can then be used to publish the full automatically generated mappings from RVK to BK, so that libraries interested in enriching their catalogs can easily access and use them.

The chosen approach to simply aggregate all classes from RVK and BK from the entries of a given cluster could also be questioned: In clusters with a large number of entries, some classes will be likely to appear more often than others, and this information is lost in the aggregation process. Future experiments should test if preserving the relative frequency of the found classes can help to improve the final mapping.

It is also planned to include additional open data sets from other libraries as sources. Dutch sources would offer the possibility of more data containing BK annotations, while other international sources could add enough DDC or LCC annotations to generate mappings between these classification systems and the RVK.

References

Aguirre, José Luis, Kai Eckert, Jérôme Euzenat and others (2012). Results of the Ontology Alignment Evaluation Initiative 2012. In Proceedings of the 7th International Workshop on Ontology Matching (OM-2012) collocated with the 11th International Semantic Web Conference (ISWC-2012), Boston, MA, USA, November 11, 2012

Aigner, Sebastian (2015). Das Informationspotential geographischer Metadaten im Kontext von Bibliotheken und webbasierten Diensten : Catalogue enrichment durch Abgleich von Metadaten am Beispiel einer Konkordanz für den Fachbereich Geographie nach Basisklassifikation (BK) und Regensburger Verbundklassifikation (RVK). Universität Wien, Master thesis.

Euzenat, Jérôme and Pavel Shvaiko (2007). Ontology matching. Heidelberg: Springer.

Isaac, Antoine, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang (2007). An empirical study of instance-based ontology matching. In Karl Aberer (Editor), The Semantic Web. 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, (Lecture Notes in Computer Science, 4825). Berlin: Springer. 253–266.

Pfeffer, Magnus (2009). Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen. In Ulrich Hohoff und Per Knudsen (Eds.), 97. Deutscher Bibliothekartag in Mannheim 2008 - Wissen bewegen, Bibliotheken in der Informationsgesellschaft. Frankfurt: Klostermann. 245–254.

Pfeffer, Magnus (2013). Using clustering across union catalogues to enrich entries with indexing information. In Myra Spiliopoulou, Lars Schmidt-Thieme, Ruth Janning (Eds.), Data analysis, machine learning and knowledge discovery. (Studies in Classification, Data Analysis, and Knowledge Organization). Berlin: Springer. 437–445

Plößnig, Veronika (2012). Konkordanzen und Kataloganreicherung in Form von Klassifikationen im Österreichischen Bibliothekenverbund (ÖBV) – ein Werkstattbericht. Retrieved September 12th, 2016, from http://epub.uni-regensburg.de/34089/1/plnig%20rvk-anwendertreffen_2012.pdf

Plößnig, Veronika; Christoph Steiner (2014). Klassifikationen: Konkordanzen, Anreicherungsprojekte und RVK - Datenkorrekturen im Österreichischen Bibliothekenverbund. Ein Update. Retrieved September 12th, 2016, from http://epub.uni-regensburg.de/34088/1/ppt%20plnig_steiner-rvk-bk-12-11-2014.pdf

Probstmeyer, Judith (2009). Analyse von maschinell generierten Korrelationen zwischen der Regensburger Verbundklassifikation (RVK) und der Schlagwortnormdatei (SWD). Hochschule der Medien Stuttgart, Bachelor thesis. Retrieved September 12th, 2016, from http://opus.bsz-bw.de/hdms/volltexte/2009/667

Lorenz, Bernd (2008). Handbuch zur Regensburger Verbundklassifikation. Materialien zur Einführung. (Beiträge zum Buch- und Bibliothekswesen, 55). Wiesbaden: Harrassowitz.

Schopman, Balthasar (2009). Instance-Based Ontology Matching by Instance Enrichment. Vrije Universiteit Amsterdam, Master thesis. Retrieved September 12th, 2016, from https://sites.google.com/site/bschopman/master-thesis

Schopman, Balthasar, Shenghui Wang, Antoine Isaac, Stefan Schlobach (2012). Instance-Based Ontology Matching by Instance Enrichment. Journal on Data Semantics, 1(4), 219–236. Retrieved September 12th, 2016, from http://link.springer.com/article/10.1007/s13740-012-0011-z

Schulz, Ursula (1991). Die niederländische Basisklassifikation: eine Alternative für die "Sachgruppen" im Fremddatenangebot der Deutschen Bibliothek. Bibliotheksdienst, 25(8), 1196–1219. Retrieved September 12th, 2016, from http://www2.bui.haw-hamburg.de/pers/ursula.schulz/publikationen/nl_bk.pdf

Waldhör, Andreas (2012). Erstellung einer Konkordanz zwischen Basisklassifikation (BK) und Regensburger Verbundklassifikation (RVK) für den Fachbereich Recht. Universität Wien, Master thesis.

Data Sharing and Identifiers

Presentation

Identifier Services: Tracking Objects and Metadata Across Time and Distributed Storage Systems

Maria Esteva
Texas Advanced Computing Center, USA
maria@tacc.utexas.edu

Ramona Walls
CyVERSE, USA
rwalls@cyverse.org

Abstract

Global identifiers are key to current and future access and reuse of data. Considering increasing data production, the complex and often messy nature of research data practices from which datasets are derived, and the ever-changing landscape of storage and publishing platforms, a single identifier type and a unique data location for a dataset does not function well nor scale. Instead, there is a need to use multiple identifiers throughout the lifecycle of a project, starting at the moment of data creation and well beyond publication to identify reuse. For complex datasets identifiers must accurately represent the diverse processes that generate the data. Thus, they must carry provenance metadata that describes these processes and make connections among their inputs and outputs. Despite the location, duplication, similarity, and archiving status of the data, its metadata must have a unified representation. These requirements have implications for implementation, including accounting for the validity of data over time, the technical resources that will support such infrastructure, and users' adoption.

Using real biology datasets, we are conducting investigations around Identifier Services (IDS). IDS is designed to bind dispersed data objects and verify aspects of their identity and integrity, independent of where the data are located and whether they are duplicate, partial, private, published, active, or static. IDS will allow individuals and repositories to manage, track, and preserve different types of identifiers and significantly improve provenance metadata of distributed collections at any point of their lifecycle.

One year into the research we have: (a) developed a generalizable data model (See figure 1) that maps genomic materials (e.g. specimen), processes (e.g. sequencing, alignment, experiments, analysis) and derived data to: global and or local identifiers and corresponding domain science (MIGS, INSDC) and citation metadata (DataCite); (b) used an API to automatically validate data associated to a global identifier and track their integrity, presence at an established location, and identity (similarity to an identical or similar dataset); and (c) implemented a user portal where the actions of the IDS are executed and its results recorded. The entities in the data model group files and metadata to corresponding processes, thus expressing their provenance. The portal provides landing pages for evolving representation of registered research projects where identifiers point to and from different data storage locations. We are using the data management infrastructure Agave (The Agave, 2016), which allows IDS to connect to repositories and access data to perform actions in a distributed computational environment. Bio-collection creators have been recruited to provide data and requirements for the prototype services, as well as structured feedback. Currently we can demonstrate a workflow in which users register their collection with IDS, select files and processes to reflect the provenance of a complex genomic dataset located both at a university storage resource and a centralized institutional repository, and conduct services across these different resources. We will report on the fitness of the data model to other science domains, provenance representation, access to the data, and the need for big data and metadata interface solutions.

Keywords: data modelling; provenance; data identifiers; distributed.

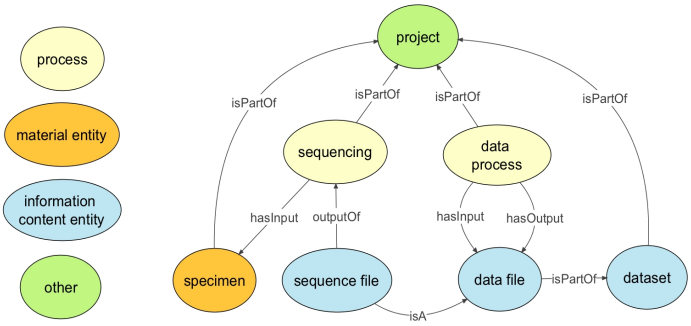


FIG. 1. Identifier Services (IDS) data model for Biology datasets adapted to Genomics.

Acknowledgements

This work is supported by the NSF EAGER: Collaborative Research: Evaluating Identifier Services for the Lifecycle of Biological Data. #26100741

References

The Agave Platform. (n.d.). Retrieved July 13, 2016, from http://agaveapi.co.

DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved January 10, 2007, from http://www.dublincore.org/documents/1998/09/dces/.

Heery, Rachel. (2004). Metadata futures: Steps toward semantic interoperability. In Diane I. Hillmann & Elaine L. Westbrooks (Eds.), Metadata in practice (pp. 257-271). Chicago: American Library Association.

Hillmann, Diane. I., Stuart A. Sutton, Jon Phipps, and Ryan J. Laundry. (2006). A metadata registry from vocabularies up: The NSDL registry project. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006, 65-75.

Lagoze, Carl, Dean Krafft, Sandy Payette, and Susan Jesuroga. (2005, November). What is a digital library anyway, anymore? Beyond search and access in the NSDL. D-Lib Magazine, 11(11). Retrieved, January 10, 2007, from http://www.dlib.org/dlib/november05/lagoze/11lagoze.html.

Presentation

Identifier Usage and Maintenance in the UNT Libraries' Digital Collections

Hannah Tarver
University of North Texas Libraries, U.S.A.
hannah.tarver@unt.edu

Mark Phillips
University of North Texas Libraries, U.S.A.
mark.phillips@unt.edu

Abstract

At the University of North Texas (UNT) Libraries we work with a large number of identifiers in relation to our Digital Collections (The Portal to Texas History, the UNT Digital Library, and the Gateway to Oklahoma History). Since our Digital Collections comprise items from other library and campus departments, as well as a large number of cultural heritage institutions across Texas and Oklahoma, many of the materials have assigned identifiers that are important to the group that owns the physical materials. We document any relevant identifiers in each item's metadata record, whether they belong to an international or established standard (e.g., ISSNs or call numbers) or have a specific context (e.g., agency-assigned report numbers).

Most discrete collections have partner-assigned identifiers that range from established accession numbers to sequentially-assigned numbers; these identifiers allow for a connection between a digital item in the public interface, copies of the associated digital files, and the physical object. To ensure that identifiers are unique within the Digital Collections, we routinely add codes that identify the partner institution at the front of each identifier, separated with an underscore (e.g., GEPFP_62-1). This makes it relatively easy to distinguish the original identifier from the code that we have added, but also prevents the inclusion of several hundred items identified as "0005" if a user wants to use an identifier to search for a particular object.

Internally, our digital infrastructure uses ARK (Archival Resource Key) identifiers to track and connect archival copies of files stored in our Coda repository with web-derivative copies in our Aubrey access system. We also currently use PURLs (Permanent Uniform Resource Locators) to identify and manage controlled vocabulary terms. For name authority, we create local authority records that act similarly to item records in terms of identifiers: each record has a system-unique identifier that generates a stable URL, but contains a field to include alternate established identifiers (e.g., ISNIs, VIAF record numbers, ORCIDs, etc.) that also refer to the entity, when applicable.

This presentation will discuss some of the complexities inherent in managing both locally-created and externally-assigned identifiers, why we use different types of identifiers throughout our infrastructure, and the implementation of various identifiers in our Digital Collections.

Presentation

Using Korean Open Government Data for Data Curation and Data Integration

Richard Smiraglia
University of Wisconsin-Milwaukee, U.S.A.
smiragli@uwm.edu

Hyoungjoo Park
University of Wisconsin-Milwaukee, U.S.A.
park32@uwm.edu

Abstract

This presentation addresses cultural heritage data-sharing practices through the use of Republic of Korea open government data for data-curation and data integration. Data curation enables data-sharing throughout the data management life cycle to create new value for new user needs. Previous studies for cultural heritage data integration have been conducted with the mediation between metadata and ontology. Examples are ontology-based metadata integration in the cultural heritage domain with mediation between Dublin Core (DC) and the meta-level ontology known as the CIDOC CRM (International Committee for Documentation - Conceptual Reference Model) (Stasinopoulou et al. 2007), DCMI type vocabulary and the CIDOC CRM (Kakali et al. 2007), DC metadata and the CIDOC CRM in cultural heritage digital object collections (Koutsomitropoulos, Solomou and Papatheodorou 2009), and between archival metadata such as Encoded Archival Description (EAD) and the CIDOC CRM (Bountouri and Gergatsoulis 2011). A gap remaining from prior studies is that cultural heritage data integration has not been actively studied with an emphasis on knowledge organization and data curation using open government data.

Our research employed a visualization phase, in which we used domain analytical techniques to better understand the contents of the population of 375 library-related open government cultural heritage data available at the Korean Open Government Website (http://data.go.kr/). Researchers translated all records from Korean to English. Data were in unstructured and in heterogeneous formats such as file formats, data formats and or web addresses.

For data curation and integration, we employed the meta-level ontology known as the CIDOC-CRM, which we applied qualitatively to small sets of carefully selected records. This phase was based on an earlier project using a different data-set (Park and Smiraglia 2014), in which cultural disparities between Korean data and the CRM were detected and resolved. Visual mappings are conducted by using the mapped Korean open government data which were in unstructured and heterogeneous formats by using CIDOC CRM version 6.2. The mappings were simple and straight-forward.

To map instantiation of records, which is required for data integration, we used FRBRoo (Functional Requirements for Bibliographic Records – object oriented), an extension of the CIDOC CRM, to map the instantiation of data records in a typical data-sharing scenario. Then, equivalent mapping processes were comparatively tested with visualizations to demonstrate the effective harmonization between the CIDOC CRM and FRBRoo, which enables the integration of metadata and data curation from unstructured and heterogeneous formats. This presentation may contribute to the cross- or meta-institutional integration of curation across institutional boundaries in cultural heritages as an imperative for cultural synergy and the role of information institutions (Smiraglia 2014) with metadata integration.

References

Bountouri, Lina, and Manolis Gergatsoulis. 2011. "The semantic mapping of archival metadata to the CIDOC CRM ontology." Journal of Archival Organization 9: 174-207.

Kakali, Constantia, Irene Lourdi, Thomais Stasinopoulou, Lina Bountouri, Christos Papatheodorou, Martin Doerr, and Manolis Gergatsoulis. 2007. "Integrating Dublin Core Metadata for cultural heritage collections Using ontologies." DCMI International Conference on Dublin Core and Metadata Applications. Singapore. 128-139. Accessed July 13, 2016. http://dcpapers.dublincore.org/index.php/pubs/article/view/871.

Koutsomitropoulos, Dimitrios A., Georgia D. Solomou, and Theodore S. Papatheodorou. 2009. "Metadata and semantics in digital object collections: A case-study on CIDOC-CRM and Dublin Core and a prototype implementation." Journal of Digital Information 10 (6).

Park, Hyoungjoo, and Richard P. Smiraglia. 2014. "Enhancing data curation of cultural heritage for information sharing: A case study using open government data." Edited by S. Closs, R. Studer, E. Garoufallou and MA Sicilia. Metadata and Semantics Research. Karlsruhe: Springer International Publishing. 95-106. Accessed July 10, 2016. doi:10.1007/978-3-319-13674-5_10.

Smiraglia, Richard P. 2014. Cultural synergy in information institutions. New York: Springer.

Stasinopoulou, Thomais, Lina Bountouri, Constantia Kakali, Irene Lourdi, Christos Papatheodorou, Martin Doerr, and Manolis Gergatsoulis. 2007. "Ontology-based metadata integration in the cultural heritage domain." Edited by Dion Hoe-Lian Goh, Tru Hoang Cao, Ingeborg Torvik Sølvberg and Edie Rasmussen. 10th International Conference on Asian Digital Libraries. Hanoi: Springer Berlin Heidelberg. 165-175. Accessed July 13, 2016. doi:10.1007/978-3-540-77094-7_25.

Posters
(Peer Reviewed)

Poster

Interoperability Workbench – Collaborative Tool for Publishing Core vocabularies and Application Profiles

Miika Alonen
CSC – IT Center for Science, Finland
Aalto University, School of Science, Finland
firstname.lastname@csc.fi

Suvi Remes
CSC – IT Center for Science, Finland
Ministry of Finance, Finland
firstname.lastname@csc.fi

Keywords: Core vocabulary; Application profile; Linked Data Modeling; Metadata repository

Introduction

The lack of semantic interoperability has been noted as an obstacle to the digital economy. As one of the solutions, the European Commission has recommended to use highly reusable metadata (EIF, 2010). In order to minimize the duplication of effort and support the interoperability in the sector of Higher Education and Research a project was established by the Ministry of Education and Culture to build a framework and tools for metadata modelling. Motivation to improve semantic interoperability comes from the need to standardize metadata management that is performed by multiple organizations in the same domain. The number of interoperability problems increase with the total number of involved parties according to Ralyté et. al. (2008). The conceptual modelling of business, services and processes, defining and maintaining terminologies, reference data and data models for multiple information systems in the same sector should no longer be seen as separate activities. The developed Semantic Information Framework (Fig. 1) describes high level architecture for linking Controlled Vocabularies, Core Vocabularies, Application Profiles and Physical Data Models. The developed framework is now also being adopted by the public administration in Finland (JHS, 2016).

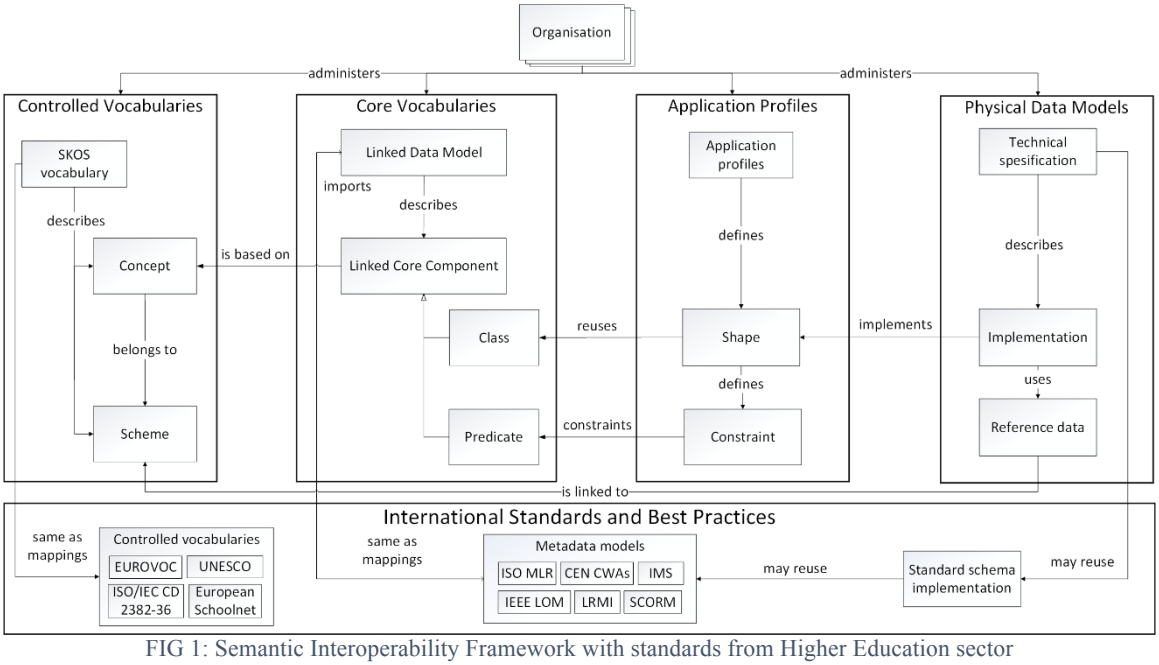


FIG 1: Semantic Interoperability Framework with standards from Higher Education sector

Semantics of information entities should be based on the terminologies that are built using systematic and formalized methods (ISO, 2009). Terminologies e.g. Vocabulary of Education (OKSA, 2016) that are typically used in interpersonal communication situations should be

published as Controlled Vocabularies in the SKOS format (SKOS, 2016) to enable them be used as a solid foundation for the semantics of the Core vocabularies and Application Profiles.

Core Vocabularies (ISA,2016) are re-usable information components that can be used to build interoperable data models. Core Vocabularies should be published as Linked Data models that are linked to the concepts in the terminology. Use of Core vocabularies and standards such as Metadata for Learning Resources (ISO/IEC 19788-1) or Metadata for Learning Opportunities (CEN/CWA 15903) should be documented as Application profiles for exposing the intended use of the metadata and to enable the measurement of the metadata quality as argued by Hillman and Phipps (2007).

Machine readable Application Profiles are used to describe data models by defining used classes, properties and constraints in RDF. However, the use of Application Profiles is not limited to documenting Linked Data models. Existing data standards and best practices may restrict the use of Linked Data in favor of other data representations. In the Semantic Interoperability Framework, we propose Application profiles to be used as technology independent documentation for all type of data representations and to create a mapping between the Universal Resource Identifiers and the local identifiers used by other type of Physical Data models.

2. Interoperability workbench

One of the challenges in reusing existing linked data models has been the lack of sophisticated tools. The envisioned synthesis of terminology work and metadata modelling also requires new workflows (Fig 2.) and tools supporting the automatic use of the controlled vocabularies.

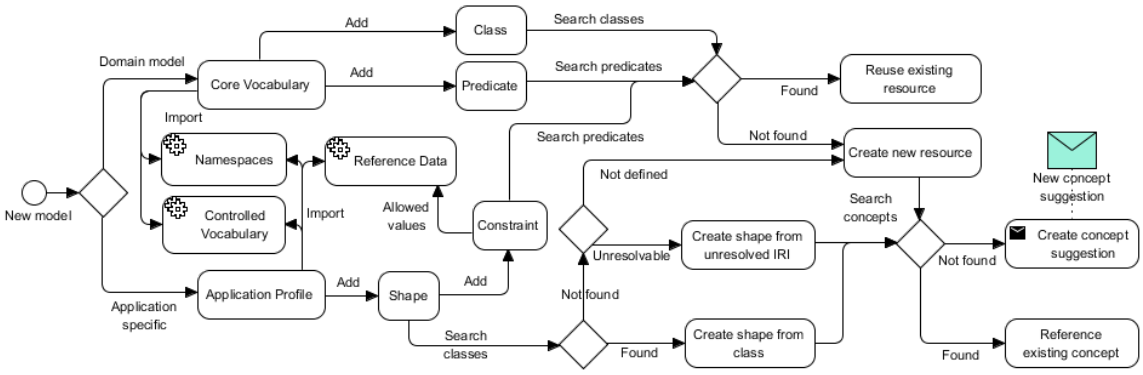


FIG 2: Simplified workflow for creating Core vocabularies and Application profiles based on shared concepts

The Interoperability Workbench (IOW, 2016) is a metadata modelling tool based on the presented workflow. The workbench is aimed for content specialists who are not experienced with RDF. The terminology used in data modelling and the data models are localized to the language preferred by the users. This allows the content specialist to collaborate with data modelers and focus on the semantics of the domain specific information structures and not the technical details of the workbench. Information structures are modeled as Core Vocabularies and Application Profiles reusing the terms and definitions from existing Controlled Vocabularies.

Core Vocabularies and Application Profiles can import existing Linked Data models by dereferencing the given namespaces. Selected controlled vocabularies are imported from the Finnish Ontology Service (Finto, 2016) and new classes and predicates are created based on the preferred terms and the definitions of the referenced concepts. Classes and predicates can also be defined based on new concept suggestions that are then forwarded to the terminology working groups. Shapes created to the Application Profiles can be based on the abstract shapes imported from the Core Vocabularies or generated from the imported Linked Data models. Shapes can also be created manually from any IRI to support the use of the unresolvable namespaces. Reference Data can also be imported to Application Profiles from various integrated sources to document the allowed values for the data.

Models are created as JSON-LD objects in JavaScript frontend and persisted to RDF database with Graph Store protocol based API. Data model of the Interoperability Workbench is documented within the workbench as an Application Profile (IOW AP, 2016). The profile extends CEN/CWA 15248 with selected SHACL (SHACL, 2016) features to support multiple class definitions and constraints. Support for additional SHACL features may also be included when needed. The Interoperability Workbench (Fig 3.) is an early prototype, but it has already been used to develop Core vocabularies in the field of Higher Education in Finland. Several application profiles also reuse the metadata models from standards and best practices (eg. EMREX, 2016 and ATT, 2016).

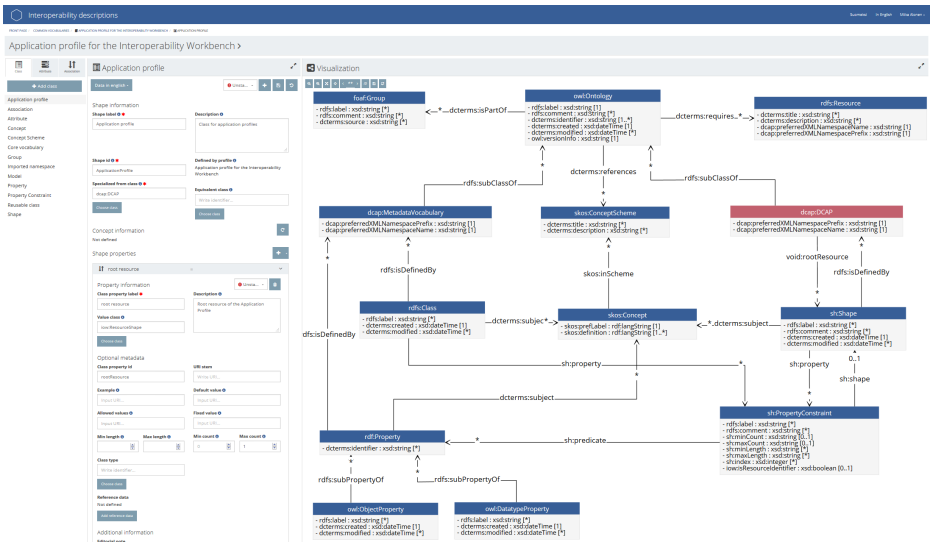


FIG 3: Screenshot of the workbench showing edit mode in Application Profile for the Workbench (IOW AP, 2016)

The Future development of the workbench will include data validation service based on the application profiles, support for URL re-direction services such as w3id.org and versioning of the data models. Development of the Interoperability Workbench and the Semantic Interoperability Framework are now putted into action as part of one of the Government Programme's key projects, *Public services will be digitalised.* A one-stop-shop service model will be developed for client-oriented usage of the key national databases and to support this, a metadata governance solution in the public administration will be implemented (Valtioneuvosto 2016*)*.

References

European Commission. (2010). European Interoperability Framework (EIF) for European public services.

Ralyte, J., Jeusfeldb, M., Backlundc, P., Kuhn, H. and Arni-Blocha, N. (2008) A knowledge-based approach to manage information systems interoperability, Information Systems, 33, 754–784.

Hillmann DI, Phipps J. Application Profiles: Exposing and Enforcing Metadata Quality. DC-2007 Conference Proceedings.

JHS, 2016, Semanttisen yhteentoimivuuden viitekehys. Retrieved May, 26, 2016 from: http://www.jhs-suositukset.fi/c/document_library/get_file?uuid=697743fc-61bd-4196-bc00-00703b65bf59&groupId=14

ISO, 2009. Terminology work -- Principles and methods (ISO 704: 2009)

OKSA, 2016. Vocabulary of Education. Retrieved from http://confluence.csc.fi/display/TIES/Sanastotyo

SKOS, 2016. SKOS Primer. Retrieved from http://www.w3.org/TR/skos-primer

ISA, 2016. Handbook for Core Vocabularies. Retrieved Aug, 10, 2016 from: https://joinup.ec.europa.eu/asset/core_vocabularies/

IOW, 2016, Interoperability Workbench. Retrieved May, 26, 2016 from: http://iow.csc.fi/

Finto, 2016, Finnish Ontology Service. Retrieved May, 26, 2016 from: http://finto.fi/

IOW AP, 2016, Retrieved May, 26, 2016 from: http://iow.csc.fi/ns/iow#

SHACL, 2016, Shape Constraint Language. Retrieved May 26. 2016 from: https://www.w3.org/TR/shacl/

CEN/CWA 15903. Metadata for Learning Opportunities

CEN/CWA 15248. Guidelines for machine-processable representation of Dublin Core Application Profiles

ISO/IEC 19788-1, MLR: Framework, Part 1. Retrievable from: http://standards.iso.org/ittf/PubliclyAvailableStandards

OILI, 2016. OILI Application Profile Retrieved May, 26, 2016 from: http://iow.csc.fi/ns/oiliu#

EMREX, 2016. EMREX Application Profile. Retrieved May, 26, 2016 from: http://iow.csc.fi/ns/emrex#

ATT, 2016. ATT Application Profile. Retrieved May, 26, 2016 from: http://iow.csc.fi/ns/att#

Valtioneuvosto, 2016. Implementation of the Government Programme. Retrieved Aug, 30, 2016 from: http://valtioneuvosto.fi/en/implementation-of-the-government-programme

Poster

Digital Asset Management Systems: Open Source or Not Open Source?

Marina Morgan
Florida Southern College,
United States
mmorgan@flsouthern.edu

Naomi Eichenlaub
Ryerson University,
Canada
neichenl@ryerson.ca

Keywords: digital asset management system; content management system; web content management system; open source; proprietary; metadata.

Abstract

The objective of this poster is to provide an overview of a number of existing open source and proprietary information management systems for digital assets. We hope that this poster will assist libraries and other institutions in their process of researching and decision-making when considering implementing a management system for their digital collections.

Background

It should be noted that while neither of the authors is currently involved in a digital asset management system migration or selection project, they have a working knowledge of all the systems described herein. Additionally, in gathering data for this project it became clear that some of the systems evaluated here have been implemented as institutional repositories as well. However, this is beyond the scope of this poster. The Digital Asset Management systems chosen for the purpose of this poster were evaluated based on their capabilities of managing a collection of digital assets such as images, videos, sound recordings, and other multimedia content.

Moreover, from the beginning it became paramount to have a clear distinction between the different terminologies used: digital asset management systems (DAM) and content management systems or web content management system (CMS). While content management systems were built to allow non-technical users to create, publish and manage website content, digital asset management systems provide an infrastructure for management and preservation of digital assets.

Introduction

As the volume of digital resources owned or created increases, many institutions want to adopt a single platform with robust functionalities for discovery, storage, and cataloging of resources. According to The National Initiative for a Networked Cultural Heritage (NINCH), "Digital Asset Management (DAM) systems provide the means to manage digital assets from creation to publication and archiving". DAM systems have become a core part of the institutions' infrastructure using rich metadata as the basis for enhanced resource discovery as well as for use in teaching and learning. These days, choosing a DAM solution invariably means choosing either an open source or a proprietary solution. Open source software has source code that is publicly available so that it can be copied, modified, and redistributed royalty-free (though usually with attribution in the form of some type of Creative Commons license) (Fitzgerald, B., Kesan, J.P. & Russo, B., 2011). The code is developed and maintained by communities of practice. Proprietary software on the other hand, is locked down in terms of access to code and made available for a fee from commercial enterprises. Unfortunately, there are no perfect products that offer off-the-shelf solutions to all the unique needs of each institution. However, there are systems that are appropriate for specific kinds of collections, as showcased below.

Methodology

For the purpose of this poster we chose three open source and three proprietary DAM systems for digital collections. They were selected based on functionality, packages and frameworks, ease of installation, number of users, scalability, metadata schemas and formats, hosting options, and technical support. Based on these characteristics, the open source DAMs reviewed are Islandora, Omeka, and DSpace, while the proprietary ones are CONTENTdm, Shared Shelf, and Digital Commons. Information was gathered by reviewing relevant literature on the topic of managing digital collections with a particular focus on collections and digital assets management systems.

1. Open Source DAMs

1.1. Islandora

a. Functionality. Born at the University of Prince Edward Island Library, Islandora is built on a software stack of FedoraCommons (repository layer), Islandora (integration layer) and Drupal (user interface layer) alongside Solr search (Ruest & Stapelfeldt, 2014). A highly extensible open-source software framework, Islandora does not have default functionality (i.e. indexing, discovery, delivery) but instead allows developers to build their own or integrate third-party options into its framework (Castagné 2013).

b. Technical Summary. Islandora uses 'solution packs' which have evolved as "best-practice workflows" from the Islandora community's experience dealing with a variety of data types (https://wiki.duraspace.org/display/ISLANDORA715/About+Islandora). Islandora excels at preserving the integrity of collections and can be customized to manage any digital asset.

c. Metadata Standards and Formats. Islandora uses MODS metadata format and generates a DC version each time the MODS is modified. Islandora generates PREMIS XML metadata on demand. You can create custom XML metadata forms as well.

1.2. Omeka

a. Functionality. Omeka is an open source web-publishing platform developed by the Roy Rosenzweig Center for History and New Media at George Mason University (https://omeka.org/). There are two options for using Omeka: you can install Omeka using LAMP (Linux, Apache, MySQL and PHP) as a self-hosted option or you can sign up for the hosted Omeka.net solution. The former requires more technical expertise and access and allows for more customization while the latter is more plug-and-play. A thorough comparison of Omeka.org vs Omeka.net is available at http://info.omeka.net/about/.

b. Technical Summary. Omeka can be populated using batch migration tools, ie. OAI-PMH (Open Archives Initiative Protocol for metadata Harvesting), CSV, EAD, or Zotero. The Omeka API allows for customizable web design and an extensive list of plugins has been developed by the Omeka community.

c. Metadata Standards and Formats. Omeka provides default metadata support for Dublin Core with a plugin for Dublin Core Extended and a METS export. Alternatively you can create your own customized metadata vocabulary.

1.3. DSpace

a. Functionality. DSpace is a cross-platform open source solution primarily used as an institutional repository platform. There are, however, a number of institutions using it as a digital asset management system (DAMs), for example the Swinburne Image Bank (http://images.swinburne.edu.au/). In terms of open source software applications for digital assets management in general, DSpace has the largest community of developers and installations and there is also now a hosted option available called DSpaceDirect.

b. Technical Summary. DSpace offers full support for OAI-PMH and SWORD (Simple Web-service Offering Repository Deposit). The latest releases of DSpace (5.x) indicate that they include support for CRUD (Create, Read, Update, Delete), linked open data, enhancements to DOI support and ORCID integration.

c. Metadata Standards and Formats. DSpace supports Qualified Dublin Core metadata with export options to many other formats including METS, MODS, RDF and MARC or you can create a custom XML metadata schema.

2. Proprietary DAMs

2.1. CONTENTdm

a. Functionality. CONTENTdm is a proprietary digital collection management system hosted and supported by OCLC. Installation is provided by OCLC, thus allowing the user to focus on creating and managing the digital collections. CONTENTdm enables a branded design and customization of the library digital collection website without prior programming skills.

b. Technical Summary. CONTENTdm has a robust technical infrastructure. It is delivered as "software as a service" (SaaS), meaning that there is no need to allocate personnel or hardware to manage the digital collections. Some of the technical features of interest are the OCR Extension to generate full-text transcripts from image files, batch importing from tab-delimited files, OAI-PMH harvesting, and operational support for incremental backups.

c. Metadata Standards and Formats. CONTENTdm can handle document, image, video and audio files of any kind. There is full control over the digital resources access, descriptions, and display. Moreover, the fully customized metadata fields maximizes user discovery of materials.

2.2. Shared Shelf

a. Functionality. Artstor's Shared Shelf Commons is an open-access library of images, a Web-based service for cataloging and managing digital collections, either as a stand-alone tool or as an add-on to the Artstor Digital Library.

b. Technical Summary. Collections are managed without local technical infrastructure or administration. They are discoverable and may be shared with other institutions or published to the Open Web via Shared Shelf Commons, the Digital Public Library of America (DPLA), or your own Omeka site. New projects can be created from existing templates or by copying a specific project and modifying the fields as needed or adding local labels. After uploading the media files, to manage the digital assets you can catalog a single record or multiple records at once using the Master Record feature. You can also export and import, update and create multiple assets using Excel.

c. Metadata Standards and Formats. Support for different media types includes video, audio, documents, and images. It also provides easy access and use of Getty vocabularies, AAT, TGN, and ULAN, which are integrated into Shared Shelf Names. Metadata templates are based on Dublin Core, VRA Core 4.0 which can be customized and extended, or samples from the Astronomy Visualization Metadata (AVM) or Darwin Core fields.

2.3. Digital Commons

a. Functionality. Digital Commons is mainly used as an institutional repository, but the Image Galleries service allows for its implementation as a digital asset management system as well. Users can use Digital Commons as a place to host all types of visual content, such as digitized archives, scanned historical documents, photographs, and other items of a visual nature.

b. Technical Summary. Digital Commons serves as an effective platform for long-term image collection preservation, enabling viewing and sharing of the collection. You can batch upload images, create dynamic slideshows, and embed them throughout the digital collection. Users can explore large, high resolution images with the use of the pan and zoom viewer that can be

displayed both on desktop and mobile devices. No longer relying on the discontinued Google API, the new content carousels and slideshows have a flexible implementation. Another consideration is that while OAI-PMH is supported to expose data, it does not however harvest OAI data from other sites.

c. Metadata Standards and Formats. Any field may be mapped to a Dublin Core value, or a custom export label. Digital Commons Image Service supports various media formats (Flash/HTML5 audio and video, Quick Time audio and video, RealAudio and RealVideo, Windows Media audio and video, YouTube, Vimeo, public domain files such as Internet Archive streaming, and other rich media via embedded API. You can add additional media fields for multiple media types.

Challenges

There are challenges with adopting both open source and proprietary software and selecting one or the other will be guided by the circumstances of each institution and even each project. In terms of implementing open source, the software may be free but there will definitely be a significant investment of staffing resources, most likely in the form of technical expertise. Alternatively, there are now a number of options to outsource open source implementation and hosting. Finally, there are well-established communities of practice to provide technical support for all the open source options described above.

On the other side, implementing proprietary solutions may be feasible for libraries without an IT staff. However, one of the major drawbacks of a proprietary-software package is expense. Depending on the number of users, the licensing and installation fee can be fairly expensive especially in comparison to open source software. While out-of-the-box solutions are easier to adopt, they are not usually as adaptable to the constantly changing needs of the institutions.

Conclusions

Choosing the right software to manage your digital collections is subjective and depends on specific circumstances, users' needs, budget, and licensing preferences. There is no shortage of options when it comes to managing, implementing, and describing your digital collections. There are proprietary systems that can be easily purchased and implemented; others require extensive knowledge of technical frameworks or a steep learning curve to implement. The software that you will choose depends on many variables and there is no perfect system when it comes to budget, needs, requirements, and implementation. Making an informed decision and evaluating all the options available will bring you closer to a system that matches the majority of your requirements.

References

bepress. (2016). Digital Commons Reference Material and User Guides. http://digitalcommons.bepress.com/reference/

Boiko, B. (2002). Content management bible. Hungry minds: New York.

Castagné, M. (2013). Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra. https://open.library.ubc.ca/cIRcle/collections/42591/items/1.0075768

CONTENTdm (2016). Support and Training. https://www.oclc.org/support/services/contentdm.en.html

Digital Asset Management and Museums - An Introduction. http://canada.pch.gc.ca/eng/1442946637162

DSpace (2016). Quickstart Guide. Retrieved May, 2016 from http://www.dspace.org/quick-start-guide

DSpace (2016). Use Case Examples: Image Repository http://www.dspace.org/repository_type/32

Fitzgerald, B., Kesan, J.P. & Russo, B. (2011). Adopting open source software: a practical guide. MIT Press.

Gkoumas, G. & Lazarinis, F. (2015). Evaluation and usage scenarios of open source digital library and collection management tools. Program: Electronic Library and Information Systems, 49(3), 226-241. doi: 10.1108/PROG-09-2014-0070

Goh, D., Chua, A., Khoo, D., Khoo, E., Mak, E. & Ng, M. (2006), A checklist for evaluating open source digital library software. Online Information Review, 30(4), 360-379. doi:10.1108/14684520610686283

Islandora (2016). Documentation http://islandora.ca/documentation

Krishnamurthy, M. (2007). Open access, open source and digital libraries: A current trend in university libraries around the world. Program: Electronic Library and Information Systems, 42(1), 48-55. doi:10.1108/00330330810851582

Larsen, R. & W. Howard. (2004) Knowledge Lost in Information: Report of the NSF Workshop on Research Directions for Digital Libraries (Chatham, Mass.: University of Pittsburg, School of Information Sciences, June 15, 2003), available online at www.digitalpreservation.gov/news/2004/knowledge_lost_report200405.pdf

NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials, Chapter XIII: Digital Asset Management. (2003) http://www.nyu.edu/its/humanities/ninchguide/XIII/

Oguz, F. (2016). Organizational Influences in Technology Adoption Decisions: A Case Study of Digital Libraries. College & Research Libraries, 77(3), 314-334. doi:10.5860/crl.77.3.314

Omeka (2016). Documentation http://omeka.org/codex/Documentation

Rath, L. (2016). Omeka.net as a librarian-led digital humanities meeting place. New Library World 117(3/4), 158-172. doi:10.1108/NLW-09-2015-0070

Reddy, R. & Wladawsky-Berger, I. (2001). Digital libraries: Universal access to human knowledge. Arlington, VA: National Coordination Office for Information Technology Research and Development.

Ruest, N. & Stapelfeldt, K. (2014). Introduction to Islandora. http://yorkspace.library.yorku.ca/xmlui/handle/10315/28006

Shared Shelf (2016). Knowledge Base http://support.sharedshelf.org/knowledge-base

Virtual Astronomy Multimedia Project. (2007). Astronomy Visualization Metadata Standard. http://www.virtualastronomy.org/avm_metadata.php

Yeh, S.T,, Reyes, F., Rynhart, J., & Bain, P. (2016). Deploying Islandora as a Digital Repository Platform: a Multifaceted Experience at the University of Denver Libraries. D-Lib Magazine, 22(7/8), doi: 10.1045/july2016-yeh

Poster

Using DC Metadata in Preservation Content: The Case of the Italian "Protocollo Informatico"

Anna Rovella
Università della Calabria
Via P. Bucci, 87036 Rende (CS), Italy
anna.rovella@unical.it

Nicola Ielpo
Università della Calabria
Via P. Bucci, 87036 Rende (CS), Italy
nicola.ielpo@unical.it

Assunta Caruso
Università della Calabria
Via P. Bucci, 87036 Rende (CS), Italy
assunta.caruso@unical.it

Keywords: Dublin Core; Metadata; Digital Preservation; Submission Information Package.

1. Introduction

As of October 2015, the digital preservation of the Protocollo Informatico (PI) by the end of the following working day is mandatory for all Italian Public Administrations. The PI is the Digital Records Management System and it plays a strategic role as regards the authenticity of the records. The inclusion of the record in the PI certifies its provenance and acquisition and determines its probative value. Starting from the PI, both embedded and external administrative work flow processes begin. Moreover, the PI register activates all the record's "properties" and "attributes" allowing for its management, such as aggregated records and its relationship with other items, its functional classification, life cycle control, appraisal and long-term preservation, access rights, processes, resources, users and roles.

2. Objectives

The purpose of this poster is to present a Metadata Element Model to support a coherent Submission Information Package (SIP) from a Records Management System to an Open Archival Information System (OAIS).

3. The Italian Digital Preservation Conceptual Model

3.1. Submission Information Packages

The Digital Preservation System must ensure the preservation (according to rules, processes and technologies) of digital information objects and it must guarantee the record's authenticity, integrity, reliability, and access. Information Packages (IP) are the preservation objects which characterize the System, and certify both processes and responsibilities. The Submission Information Package is the information package that the records creator sends to the Digital Preservation System. The strategies used in the creation of a SIP are fundamental in order to coherently transfer objects and information from the PI to the Digital Preservation System. It is clear that the Submission Information Package contains the records and metadata appropriately linked to processes and functional reference models and that SIP quality is closely related to the quality of the Digital Preservation System.

3.2. Metadata

The Submission Package is made up of one or more digital objects and of metadata which permits representation and access over time in a Digital Preservation Ecosystem. This Ecosystem is populated by various stakeholders, each with different responsibilities. The relationship between the information object and metadata allows the Information Package to represent the relationship amongst the objects along with the entities of the environment. The Information Package is characterized by:

a) Content Information;

b) Preservation Description Information (PDI, *i.e.* Reference Information, Context Information, Provenance Information, Fixity Information and Access Rights Information).

Preservation Systems help repositories manage diverse metadata and facilitate the exchange of metadata or Information Packages between repositories. Metadata quality is one of the key elements towards the successful application of the System.

3.2.1. Using Dublin Core Metadata

In the light of what has been expressed thus far, we decided to develop a Metadata Element Model using the Dublin Core as a base. This model maps metadata elements to Dublin Core qualified terms, conceived as the backbone of efficiency and as a harmonizing interchangeable bridge between various identifying ways to manage and preserve digital objects and records from various domains.

The choice of the Dublin Core was prompted by its characteristics of simplicity, widespread semantic interoperability towards the preservation repository and metadata crosswalk into other repositories. Moreover, the Italian regulations regarding digital preservation recommend the use of ISO 15836:2003 and the Protocollo Informatico uses the DC Metadata to support records registration. The proposed Metadata Schema extends the Dublin Core also in accordance with the Singapore Framework for Dublin Core Application Profiles. The schema uses logic and structuring which are also typical of other metadata schemas such as ISO2308-1-2:2009, PREMIS (2015), METS, MODS, etc.

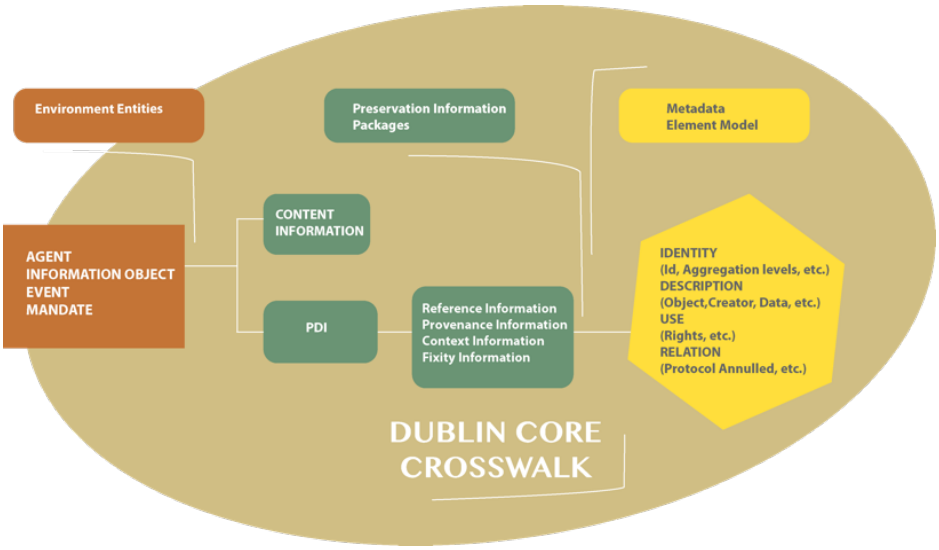


FIG. 1: Conceptual Model.

The graphical representation in (FIG. 1) illustrates the Conceptual Model on which the Metadata Element Model was based and how it interacts with the Dublin Core. The overall structure of the Conceptual Schema shows how metadata represent the Environment Entities and their relationships in Information Packages and characterize the Digital Information Object life cycle within different processes and workflows.

3.2.2. Development of the model

The designed Metadata Element Model consists of two schemas: the first defines the metadata for the administrative record or more generally speaking for information objects (e.g. TABLE 1), the second for the file and aggregated records. Each schema is set up in a hierarchical structure and has variable granularity and extensions by way of authority lists, controlled vocabularies, etc., making the metadata schema sufficiently rich, not only in the element number, but also consistent

enough to describe different characteristics of heterogeneous information objects. In the development of the Model, the combination and use of metadata schemas as well as an analysis of possible problems (incorrect values, incorrect elements, missing information, information loss, inconsistent value representation) have been taken into consideration. The model has been used and evaluated by several domain experts from different Italian regions in the context of a specific agreement between CNR, University of Calabria and ItConsult S.p.a. At the time of writing, the Agency for Digital Italy (AgID) is evaluating the model for its integration into government policy documents. We expect the results of this testing phase to properly evaluate the application and work out any critical issues.

TABLE 1: Example of some elements in Administrative Record Metadata Schema.

ELEMENT	ALLOWED VALUES		DESCRIPTION	CROSSWALK
ID	Alphanumeric string		An unambiguous and persistent reference to the digital information object within a given context.	<dcterms:Identifier>
Creator	Name: String Surname: String CodiceFiscale: Alphanumeric string	IPA Code	Compound metadata element for identifying the entity primarily responsible for making the resource. (DCMI Metadata Terms, 2012).	<dcterms:Creator> <dcterms:Identifier>
	Surname: String CodiceFiscale: Alphanumeric string			
	Surname: String CodiceFiscale: Alphanumeric string			
Rights	RightsType: string		Compound metadata element that defines the type and validity of rights and permissions on record. Possibly associated with controlled list.	<dcterms: RightsType >
	RightsDate: date and time			<dcterms: RightsDate >
	RightsHolder: name, surname, CF, IPA.			<dcterms: RightsHolder>
Timestamp/ Inscription	Data and time		Date and time of record production (UTC).	<dcterms:DateValid>

4. Conclusion and future work

This poster introduces research aimed to design an extensible Metadata Element Model for content preservation within the context of Italian digital administrative records. The goal of the project is to develop, test and promote a standard interchange format for exchanging stored information packages among OAIS-based preservation repositories. In the future, we plan to work on the semantic level by optimizing authority control, with the definition of authority lists for the core elements and to enhance the use of standard vocabularies and make them compatible with the international standards.

References

AgID, Produzione e Conservazione del Registro Giornaliero di Protocollo, 01/10/2015. http://www.agid.gov.it/sites/default/files/documenti_indirizzo/produzione_e_conservazione_del_registro_giornaliero_di_protocollo_0.pdf and Linee Guida sulla Conservazione dei Documenti Informatici, 10/12/2015. http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def_.pdf

Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

The Singapore Framework for Dublin Core Application Profiles. http://dublincore.org/documents/singapore-framework/.

ISO 23081-2:2009. Information and documentation - Managing metadata for records - Part 2: Conceptual and implementation issues.

DCMI Metadata Terms, 2012, http://dublincore.org/documents/dcmi-terms/.

PREMIS Data Dictionary for Preservation Metadata, Version 3.0, June 2015. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

METS Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/.

MODS Metadata Object Description Schema. http://www.loc.gov/standards/mods/.

Poster

Modeling Cultural Evolution with Metadata Collections: A Phylomemetic Approach

Nicholas M. Weber
University of Washington, USA
nmweber@uw.edu

Andrea K. Thomer
University of Illinois at Urbana-Champaign, USA
thomer2@illinois.edu

Keywords: Evolution of metadata; collections; formal modeling; phylomemetics

1. Introduction

Descriptive metadata is typically used to record information about digital artifacts and thereby facilitate users' retrieval and use of these artifacts. The resulting collections of descriptive metadata records may be considered digital artifacts in and of themselves, and evidence of the behavior and values of the communities and cultures that produce, use, and cooperate in provisioning digital artifacts. Studying the ways that collections of metadata records change over time may reveal novel insights into the evolution of the communities that not only create digital artifacts – but that catalog and manage them as well.

In this poster we describe and apply an approach to modeling cultural evolution, *phylomemetic analysis,* using collections of metadata records. We show that collections of descriptive metadata records can be used as a primary data source for the evolutionary analysis of institutions and communities engaged in digital scholarship, and discuss the potential implications of this method for metadata repository managers and researchers alike.

2. Phylomemetics

Derived from (and named after) phylogenetic methods in evolutionary biology, phylomemetics refers to the evolutionary analysis of non-genetic or biological data (Howe & Windram, 2011). In a phylogenetic analysis of biological specimens unique aspects of an organism (e.g. DNA sequences, the number of toes on a limb; the presence or absence of a hair or feathers; or as Darwin himself demonstrated, the different shapes of birds' beaks) are coded qualitatively as *characters* and then statistically analyzed to infer an evolutionary tree. In a phylomemetic analysis, "memes" rather than genes are coded and analyzed. This approach has previously been used to study cultural evolution through a range of artifacts, both physical and conceptual (e.g. cornets (Tëmkin & Eldredge, 2007), arrowheads (O'Brien, Darwent & Lyman 2001), languages (Bates and Elman, 2000), music (Le Bomin, Lecointre & Heyer, 2016), and folk tales (Tehrani, 2013)).

Just as descriptive metadata from digital libraries, such as the HathiTrust, can be studied through "distant readings" of cultural trends over time (Underwood, 2016), so can collections of metadata records. For example, in previous work we've shown that phylomemetic methods can be applied to collections of NASA metadata records by using attribute-value pairs as characters; in doing so, we are able to identify clusters of user communities altering an earth science dataset for similar purposes, as well as points at which communities split apart from one another (Thomer and Weber, 2014). Thus, descriptive metadata collections can be used to model cultural change within communities that produce, share, and alter datasets and other digital artifacts. Though this change is often self-reported to a degree through texts such as journal articles, software notes, or even the "about" pages of an organization's website, the phylomemetic approach provides an alternative line of evidence to support – or challenge – existing narratives of a community's history. Additionally, understanding how the content or completeness of

metadata records evolve over time can inform the work of metadata creators, and metadata repository managers. For instance, changes in how users create records (e.g. filling in more or less fields, with more or less clarity) can be indicative of larger trends within a community that may need to be addressed by alterations in policy or best practices. A phylomemetic view of metadata collections may help repository managers understand and guide their user communities.

3. Software Package Metadata

Here we demonstrate this approach with an analysis of metadata records describing different packages of the Debian operating system. The Debian operating system is one of the most successful distributions of Linux – a free open source software alternative to commercial operating systems, such as Windows and Mac OS. Each new distribution of Debian contains over four hundred individual different software packages. For instance, just as each distribution of Windows comes with a word processing package (e.g. Microsoft® Word) so too does Debian (e.g. AbiWord). The different package configurations of a distribution represent significant changes in the people and the politics of an open-source project as an institution. While these changes are described in the software documentation, our phylomemetic analysis will provide us with an alternative line of evidence, through which we may better understand the changes of this software, and its development community, over time.

The workflow we have developed is as follows:

- We harvest descriptive metadata records about different software packages found in each Debian distribution (e.g. word processing software packages). Each package's metadata are coded to create a character matrix.
- A package's character matrix represents differences or changes in a package over time - collectively the different package matrixes represent the 'genetic makeup' of a Debian distribution. This is much like a biologist would compare individual characteristics of one specimen to another and code for absence or presence of common features.
- We then load this matrix into phylogenetic software - PAUP (Swofford, D. L., & Begle, 2013) - to produce a visualization of the different Debian distributions.
- We set PAUP to use a maximum likelihood algorithm - which sorts characteristics by their relevant distance (difference) from one another.
- PAUP then produces a tree' that visualizes the relevant divergence of each distribution.
- The tree can then be used to infer differences in how package configurations represent differences in Debian distributions, potentially revealing substantive changes in the institutional features of the broader open-source project.

4. Future Work

Phylomemetic studies of metadata collections are related to a number of previous evolutionary studies in knowledge representation and classification research. For instance, work by Krause et al (2015) and Tennis (2002, 2012) is of particular relevance to the modeling of cultural change using metadata as a primary source. We follow these authors in noting that metadata creation methods evolve just as much as the artifacts they describe; thus, a phylomemetic analysis is potentially a way to not just study the relatedness and evolution of records, but the evolution of different methods of metadata application development and design.

References

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology*, *54*(4), 513-526.

Howe CJ, Windram HF (2011) Phylomemetics—Evolutionary Analysis beyond the Gene. PLoS Biol 9(5): e1001069. doi: 10.1371/journal.pbio.1001069

Krause, E. M., Clary, E., & Greenberg, J. (2015). Evolution of an Application Profile: Advancing Metadata Best Practices through the Dryad Data Repository. In International Conference on Dublin Core and Metadata Applications (pp. 63-75).

Le Bomin, S., Lecointre, G., & Heyer, E. (2016). The Evolution of Musical Diversity: The Key Role of Vertical Transmission. *PloS one*, *11*(3), e0151570.

O'Brien, M. J., Darwent, J., & Lyman, R. L. (2001). Cladistics Is Useful for Reconstructing Archaeological Phylogenies: Palaeoindian Points from the Southeastern United States. *Journal of Archaeological Science*, *28*(10), 1115–1136. doi:10.1006/jasc.2001.0681

Swofford, D. L., & Begle, D. P. (1993). *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1, March 1993*. Center for Biodiversity, Illinois Natural History Survey.

Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. Current Anthropology, 48(1), 146-154.

Tennis, J. T. (2002). López-Huertas, M. (ed.) Subject Ontogeny: Subject Access through Time and the Dimensionality of Classification. In Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference. Vol. 8. 54 - 59. Ergon Verlag. Würzburg.

Tennis, J. T. (2012). The strange case of eugenics: A subject's ontogeny in a long-lived classification scheme and the question of collocative integrity. Journal of the American Society for Information Science and Technology, 63(7), 1350-1359.

Tehrani, J. J. (2013). The phylogeny of little red riding hood. *PloS one*, *8*(11), e78871.

Thomer, A. K., & Weber, N. M. (2014). The phylogeny of a dataset. Proceedings of the American Society for Information Science and Technology, 51(1), 1-11.

Underwood, T. (2016) The Lifecycle of Genres. Journal of Cultural Analytics. 1(1). Retrieved from: http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/

Dolmen: A Linked Open Data Model to Enhance Museum Object Descriptions

Clément Arsenault
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada
clement.arsenault@umontreal.ca

Elaine Ménard
School of Information Studies, McGill University, Canada
elaine.menard@mcgill.ca

Abstract

This paper presents the DOLMEN project (Linked Open Data: Museums and Digital Environment), offering to develop a linked open data model that will allow Canadian museums to disseminate the rich and sophisticated content emanating from their various databases and to, in turn, make their cultural and heritage collections more accessible to future generations. The rationale, specific objectives, proposed methodology and expected benefits are briefly presented and explained.

Keywords: Linked Open Data; Museums; Canada

1. The DOLMEN Project

1.1. Rationale

Despite the latest technological advancements, the possibility for museums to provide access to their collections via the web remains a pressing concern for most. Many factors can explain this situation, the main one being incompatibility of data formats among museums. Cultural institutions often work in silos and do not use standardized description schemes. This lack of interoperability results in the near impossibility to exchange data among museums, therefore multiplying the colossal task of producing descriptions for the multitude of artifacts in their collection. The description of a museum object (e. g., a famous painting) in a museum database will usually include a restricted selection of information elements such as a simple photograph of the painting, the artist's name, the year of creation, the dimensions, the techniques used, and other basic descriptive metadata. In addition, specific managerial metadata (e. g., acquisition number, condition reports, storage notes, handling and manipulation of objects, information on crating, etc.), often judged irrelevant for the public, may not be displayed to the community at large.

Our research project, DOLMEN (Linked Open Data: Museums and Digital Environment), offers to develop a linked open data model that will allow Canadian museums to disseminate the rich and sophisticated content emanating from their various databases and to, in turn, make their cultural and heritage collections more accessible to future generations. A few linked open data projects, focusing specifically on museum objects, have recently been launched (Oard et al., 2014). However, these projects are not yet widespread, and most Canadian museums still hesitate to embrace that model. Given a worrying lack of financial resources, Canadian museums often feel helpless in the face of fast-paced technological evolution. This illustrates the pressing need to conduct extensive research on linked open data. The desire to transmit and share digital content requires museums to integrate a collaborative work logic, both among themselves and with other data providers. Making use of linked open data will answer three specific needs for museums: speeding up processes, gaining visibility and reducing costs. With the unprecedented potential of the semantic web and collaboration between researchers from different disciplines, the DOLMEN project will allow museums to offer expanded access to the descriptive multilingual content

associated with their digital collections. In turn, this will allow them to address a broader public, which is, for most museums, a fundamental mission.

1.2. Objectives of the Project

DOLMEN offers to examine the fundamental elements for the description of museum objects and model them by using linked open data. More specifically, three objectives have been established: (1) Theoretical: To understand the characteristics necessary for the description of museum objects of any kind; (2) Empirical: To define a model for the description of museum objects using linked open data; and (3) Practical: To strengthen data exchange networks among various cultural and heritage institutions. The DOLMEN project is a stepping stone towards implementing the semantic web, as envisioned by Berners-Lee, Hendler and Lassila (2001) more than a decade ago, with the aim of making cultural and heritage collections more accessible to future generations.

1.3. Proposed Methodology

The proposed methodology for the DOLMEN project comprises three phases. For the first phase of the research project, we will examine a sample of databases from Canadian museums of different types and sizes; a sample of 150 to 200 databases is considered. The museums will be selected to cover a wide array of museum objects that will eventually be described with DOLMEN.

Phase I of the project will start during fall 2016, will begin with an exhaustive inventory of open terminology databases, in English or French (Canada's official languages), that can be used for the description of museum objects. Non-textual databases will also be included and will come from all types of cultural heritage organizations. This survey will also be looking at existing descriptive standards, models and schemas (CIDOC-CRM, LIDO, CDWA-lite, EDM, etc.) in order to assess their suitability for the project. An analysis of large aggregation projects that are already in place (e.g., Athena, Smithsonian American Art Museum, Proof-of-Concept, etc.) will also serve as guidance to refine our methodology for the subsequent phases of the study. The results of this first phase of the study will provide the foundation of the DOLMEN linked open data model and guide us on a number of relevant issues, such as, differences in metadata formats, terminological aspects, cost and feasibility, and reliability of data providers.

Phase II of the methodology proposes model structuring using the descriptive elements and open data content identified in Phase I. This modelling comprises three main steps: (1) encoding of descriptive elements with the Resource Description Framework (RDF); (2) creation of links between metadata converted into RDF and open data sources, and other repositories; and finally (3) validation of open data links to ensure that data is accurate and that links to other open sources are properly accessible.

Finally, Phase III of the research project will involve the evaluation of DOLMEN. The assessment will focus on the linked open data obtained to estimate the completeness and specificity level of the model. This will be achieved by asking a sample of approximately 150 participants to examine and assess the data provided by DOLMEN which will in turn allow us to measure the degree of effectiveness and efficiency of the model, and to survey the participant satisfaction regarding the informational content offered by the model.

1.4. Expected benefits of the Project

The possibility to create links between different databases offers a wide range of possibilities to cultural institutions. The use of linked open data creates a new context for enriching museum objects descriptions within existing metadata records and linking them to semantically related resources. In other words, object descriptions will be improved by adding data provided by various museums and other cultural resources databases. DOLMEN is intended to be an innovative tool for both professionals working in museums and the general public. With the

integration of text and multimedia content (e. g., 3D images, sound recordings, etc.) this will constitute a benefit for users with specific information needs. DOLMEN is leading the way to provide better access to Canadian cultural and heritage collections through linked open data. More specifically, linked open data will enable web users and third party organizations to integrate resources to create richer, more sophisticated and more interoperable metadata for museum objects.

Acknowledgements

The authors wish to acknowledge the contribution of the Canadian Social Sciences and Humanities Research Council for supporting this research project (grant number 435-2016-0460).

References

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001, May). The semantic Web. *Scientific American 284*(5): 34–43. doi: 10.1038/scientificamerican0501-34.

Oard, Douglas W., Amalia Levi, Ricardo Punzalan, and Rob Warren (2014, April). Bridging communities of practice: Emerging technologies for content-centered linking. Paper presented at "MW2014," the Annual Conference of Museums and the Web, Baltimore, MD, April 2014. Retrieved from http://mw2014.museumsandtheweb.com/paper/bridging-communities-of-practice-emerging-technologies-for-content-centered-linking/.

Posters
(Best Practice)

AUTHOR INDEX

METADATA SUMMIT