Metadata for Semantic and Social Applications

Edited by Jane Greenberg
and Wolfgang Klas

DC-2008 BERLIN

Proceedings of the International Conference on Dublin Core and Metadata Applications

22–26 September 2008

Universitätsverlag Göttingen

Jane Greenberg, Wolfgang Klas (Ed.)

Metadata for Semantic and Social Applications

Published by Universitätsverlag Göttingen 2008

Metadata for Semantic and Social Applications

Proceedings of the International Conference
on Dublin Core and Metadata Applications
Berlin, 22-26 September 2008

DC 2008: Berlin, Germany

Edited by Jane Greenberg and Wolfgang Klas

Published by the Dublin Core Metadata Initiative,
Singapore

and

Universitätsverlag Göttingen
2008

Bibliographical information of the German National Library

This publication is recorded by the German National Library; detailed bibliographical data are available here:
<http://dnb.ddb.de>

Typesetting	Beate Rabold, Sandra Lechelt, Susanne Dobratz Humboldt- Universität zu Berlin
Cover Design	Mirjam Kessler, German National Library Manuela Schulze, Humboldt-Universität zu Berlin Margo Bargheer, Universitätsverlag Göttingen
Production Managament	Stefanie Rühle Niedersächs. Staats- und Universitätsbibliothek Göttingen

http://univerlag.uni-goettingen.de
ISBN: 978-3-940344-49-6
ISSN: 1939-1358

Preface

Establishing a standard like Dublin Core is like building a bridge. It makes exchange possible. Fittingly, the Dublin Core conference 2008 is taking place in Berlin, which is often called a bridge between Western and Eastern Europe. This event reaches even beyond Europe, with registered participants from nations all over the world such as United States, South Africa, Japan and New Zealand.

This year's conference will focus on metadata for social and semantic applications. For the first time, alongside the English tutorials there will be tutorials in German, prepared and presented by students from the University of Applied Sciences Potsdam. After three days of plenary as well as parallel sessions the conference will close on Friday with four different seminars focussing on interoperability and metadata vocabularies.

Organizing such a conference would be impossible without the invaluable help of the following six organizations: the Competence Centre for Interoperable Metadata (KIM), the Max Planck Digital Library (MPDL), the Göttingen State and University Library (SUB), the German National Library (DNB), Humboldt-Universität zu Berlin (HU Berlin) and the Dublin Core Metadata Initiative (DCMI).

For the funding we would like to thank the German Research Foundation (DFG) and the Federal Ministry of Education and Research (BMBF). Additionally we would like to thank Elsevier, the Common Library Network GBV, IBM, OCLC and Sun Microsystems for their generous sponsoring.

Last but not least, the conference is supported by Wikimedia Deutschland, local support community of Wikipedia, the well-known online encyclopedia.

We are sure that this year's conference will serve as a bridge between the participants and their knowledge, ideas and visions.

With sincere wishes for a productive conference,

Heike Neuroth
Niedersächsische Staats- und Universitätsbibliothek Göttingen
on behalf of the DC-2008 Host Organisation Committee

Preface

It is with great pleasure that DCMI welcomes participants to DC-2008, the 8th annual International Conference on Dublin Core and Metadata Applications to Berlin, Germany.

For DCMI, it is also a return to Germany, as we came to Frankfurt in 1999 for one of the last invitational workshops, before the conference cycle began in 2001.

A lot has happened since then, most notably the establishment of the Dublin Core Metadata Element Set as ISO standard 15836. Important steps since then include the development of the extended set of DCMI Terms, the DCMI Abstract Model and the Singapore Framework for Dublin Core Application Profiles.

Since those days in 1999, the Dublin Core community has grown from a small and committed group of metadata pioneers into a large community of researchers and practitioners, who come together once a year to share experiences, discuss common issues and meet people from across the planet.

This year in Berlin, the program has a dual focus, with attention for semantic applications (where the focus is on machine-readable information and co-operation between automated systems) and social applications (where the focus is on co-operation between people). We believe that both forms of co-operation are crucial for enabling the interoperability that is at the heart of our work on Dublin Core metadata.

As usual, we hope that the event in Berlin will help people to gain understanding of approaches and developments in many places around the world, in many application domains and in many languages, and at the same time allow participants to get to know each other and build and extend personal and professional networks.

On behalf of DCMI and its many contributors, I would like to wish everybody a very useful and pleasant conference.

Makx Dekkers
Managing Director
Dublin Core Metadata Initiative

Acknowledgements

The DC2008 proceedings represented in the following pages are the end-result of a long process that includes the submission of papers, reports, and posters; reviewing the submissions; organizing the accepted work among themes; and reviewing and formatting the final copies of the accepted works for publication in these proceedings. The process required the input of all members of the Program Committee and the Publications Committee. As conference Co-Chairs, we'd like to thank and acknowledge the efforts of the members of both of these committees, and, in particular, the outstanding and tremendous efforts of Stuart Sutton, Hollie White, Bernhard Haslhofer, Stefanie Rühle, and Susanne Dobratz.

Jane Greenberg, University of North Carolina at Chapel Hill
Wolfgang Klas, Universität Wien
Program Committee Co-Chairs, DC-2008

Introduction

The 2008 International Conference on Dublin Core and Metadata Applications (DC-2008) is the sixteenth Dublin Core workshop, and the eighth full conference program to include peer-reviewed scholarly works (Tokyo, 2001; Florence, 2002; Seattle, 2003; Shanghai, 2004; Madrid, 2005; Manzanillo, 2006; and Singapore, 2007).

DC-2008 takes place in Berlin, Germany, a vibrant city in which cultural and scientific ideas are exchanged daily among the many sectors of society. Home to some of the world's most significant libraries and scientific research centers, Berlin is an ideal location for DC-2008, and for further linking the community of researchers, information professionals, and citizens who increasingly work with metadata to support the preservation, discovery, access, use, and re-use of digital information and information associated with physical artifacts.

The theme for DC-2008 is "Metadata for Semantic and Social Applications". Standardized, schema-driven-metadata underlies digital libraries, data repositories, and semantic applications leading toward the Semantic Web. Metadata is also part of the fabric of social computing, which includes the use of wikis, blogs, and tagging. These two trends flow together in applications such as Wikipedia, where authors collectively create structured information that can be extracted and used to enhance access to and use of information sources.

The papers in these proceedings address an array of significant metadata issues and questions related to metadata for semantic and social applications. The proceedings include twelve papers that are organized among the following five themes: 1. Dublin Core: Innovation and Moving Forward; 2. Semantic Integration, Linking, and KOS Methods; 3. Metadata Generation: Methods, Profiles, and Models; 4. Metadata Quality; and 5. Tagging and Metadata for Social Networking. The proceedings also include eight reports distributed among the following three themes: 1. Toward the Semantic Web, 2. Metadata Scheme Design, Application, and Use; and 3. Vocabulary Integration and Interoperability. The last part of the proceedings includes twelve extended one-page abstracts capturing key aspects of current research activities.

These papers, reports, and poster abstracts present a cross-section of developments in the field of metadata, with particular attention given to several of the most pressing challenges and important successes in the area of semantic and social systems. Their publication serves as a record of the times and provides a permanent body of knowledge upon which we can build over time.

We are pleased to have representation of such high quality work and to have had the input and review of an outstanding Program Committee in making the selection for this year's conference. We are also pleased that the DC-2008 is taking place in Berlin, a city of international culture. Finally, we are honored to have had the opportunity to serve as this year's Program Committee Co-Chairs and bring you a fine collection of work from our colleagues around the world.

Jane Greenberg, University of North Carolina at Chapel Hill
Wolfgang Klas, Universität Wien
Program Committee Co-Chairs, 2008

Conference Organization

Conference Coordinators

Makx Dekkers, Dublin Core Metadata Initiative
Heike Neuroth, Göttingen State University Library/ Max Planck Digital Library Germany

Program Committee Co-Chairs

Jane Greenberg, School of Information and Library Science, University of North Carolina, USA/ SILS Metadata Research Center
Wolfgang Klas, Multimedia Information Systems Group, University of Vienna, Austria

Program Committee

Abdus Sattar Chaudhy, Nanyang Technological University, Singapore
Aida Slavic, UDC Consortium, The Hague, The Netherlands
Alistair Miles Rutherford, Appleton Laboratory, UK
Allyson Carlyle, University of Washington, USA
Ana Alice, Baptista University of Minho, Portugal
Andrew Wilson, National Archives of Australia, AUS
Andy Powell, Eduserv Foundation, UK
Ann Apps, The University of Manchester, UK
Bernhard Haslhofer, University of Vienna, Austria
Bernhard Schandl, University of Vienna, Austria
Bradley Paul Allen, Siderean Software, Inc., USA
Charles McCathie, Nevile Opera, Norway
Chris Bizer, Freie Universität Berlin (FU Berlin), Germany
Chris Khoo, Nanyang Technological University, Singapore
Cristina Pattuelli, Pratt Institute, USA
Corey Harper, New York University, USA
Dean Kraft, Cornell University Library, USA
Diane Ileana Hillmann, Cornell University Library, USA
Dion Goh, Nanyang Technological University, Singapore
Eric Childress, OCLC, USA
Erik Duval, Dept. Computerwetenschappen, Katholieke Universiteit Leuven, Belgium
Eva Mendez, University Carlos III of Madrid, Spain
Filiberto F. Martinez-Arellano, National Autonomous University of Mexico, Mexico
Gail Hodge, Information International Associates, USA
Hollie White, University of North Carolina, Chapel Hill, USA
Igor Perisic, LinkedIn, USA
Jacques Ducloy, Institut de l'Information Scientifique et Technique, France
Jane Hunter, University of Queensland, AUS
Jian Qin, Syracuse University, USA
John Kunze; California Digital Library, USA
Joseph A. Busch, Taxonomy Strategies LLC, USA
Joseph Tennis, University of Washington, USA
Juha Hakala, Helsinki University Library - The National Library of Finland, Finland
Kathy Wisser, University of North Carolina, Chapel Hill, USA
Leif Andresen, Danish Library Agency, Denmark
Liddy Nevile, Dep. of Computer Science & Computer Engineering, La Trobe University, AUS
Marcy Lei Zeng, Kent State University, USA
Michael Crandall, University of Washington, USA

Miguel-Angel Sicilla, University of Alcala, Spain
Mikael Nilsson, Royal Institute of Technology, Sweden
Mitsuharu Nagamori, University of Tsukuba, Japan
Paul Miller, Talis, UK
Pete Johnston, Eduserv Foundation, UK
Sandy Roe, Illinois State University, USA
Sebastian R. Kruk, DERI Galway, Ireland
Shigeo Sugimoto, University of Tsukuba, Japan
Stefanie Rühle, Research & Development Department, Göttingen State University Library, GERMANY
Stuart A. Sutton, University of Washington, USA
Stuart Weibel, OCLC, USA
Tom Baker, Research & Development Department, Göttingen State University Library, Germany
Traugott Koch, Max Planck Digital Library, Germany
Wei Liu, Shanghai Library, China
William E. Moen, University of North Texas (UNT), USA

Workshop Committee

Makx Dekkers, Dublin Core Metadata Initiative

Tutorial Committee

Achim Oßwald, Institute of Information Science, University of Applied Sciences, Germany
John Roberts, Archives Management, New Zealand

Publications Committee

Susanne Dobratz, (Co-Chair), Humboldt-Universität zu Berlin, Germany
Stefanie Rühle, (Co-Chair), Göttingen State University Library, Germany
Patrick Danowski, WikiMedia, Germany
Sandra Lechelt, Humboldt-Universität zu Berlin, Germany
Beate Rabold, Humboldt-Universität zu Berlin, Germany
Hollie White, University of North Carolina, Chapel Hill, USA

Publicity Committee

Mirjam Keßler (Chair), German National Library Frankfurt, Germany
Monika Nisslein, Max Planck Digital Library, Germany
Cornelia Reindl, Max Planck Digital Library, Germany

Host Organisation Committee

Laurent Romary, Max Planck Digital Library, Germany
Michael Seadle, Humboldt-Universität zu Berlin, Germany
Reinhard Altenhöner, German National Library Frankfurt, Germany
Stefanie Rühle, Göttingen State University Library, Germany

assisted by

Peggy Beßler; Malte Dreyer; Susanne Dobratz; Stefan Farrenkopf; Christine Frodl; Katrin Gashi; Elke Greifeneder; Justine Haeberli; Rupert Kiefl; Traugott Koch; Matthias Schulz; Ulla Tschida; Andre Wobst

CONTENTS

PAPER SESSION 1 DUBLIN CORE: INNOVATION AND MOVING FORWARD

PAPER SESSION 2 SEMANTIC INTEGRATION, LINKING, AND KOS METHODS

PAPER SESSION 3 METADATA GENERATION: METHODS, PROFILES, AND MODELS

PAPER SESSION 4 METADATA QUALITY

PAPER SESSION 5 TAGGING AND METADATA FOR SOCIAL NETWORKING

PROJECT REPORT SESSION 1 TOWARD THE SEMANTIC WEB

PROJECT REPORT SESSION 2 METADATA SCHEME DESIGN, APPLICATION, AND USE

PROJECT REPORT SESSION 3 VOCABULARY INTEGRATION AND INTEROPERABILITY

POSTER ABSTRACTS

Full Papers

Session 1:

Dublin Core: Innovation and Moving Forward

Encoding Application Profiles in a Computational Model of the Crosswalk

Carol Jean Godby
OCLC, USA
godby@oclc.org

Devon Smith
OCLC, USA
smithde@oclc.org

Eric Childress
OCLC, USA
childress@oclc.org

Abstract

OCLC's Crosswalk Web Service (Godby, Smith and Childress, 2008) formalizes the notion of *crosswalk*, as defined in Gill,et al. (n.d.), by hiding technical details and permitting the semantic equivalences to emerge as the centerpiece. One outcome is that metadata experts, who are typically not programmers, can enter the translation logic into a spreadsheet that can be automatically converted into executable code. In this paper, we describe the implementation of the Dublin Core Terms application profile in the management of crosswalks involving MARC. A crosswalk that encodes an application profile extends the typical format with two columns: one that annotates the namespace to which an element belongs, and one that annotates a 'broader-narrower' relation between a pair of elements, such as Dublin Core *coverage* and Dublin Core Terms *spatial*. This information is sufficient to produce scripts written in OCLC's Semantic Equivalence Expression Language (or Seel), which are called from the Crosswalk Web Service to generate production-grade translations. With its focus on elements that can be mixed, matched, added, and redefined, the application profile (Heery and Patel, 2000) is a natural fit with the translation model of the Crosswalk Web Service, which attempts to achieve interoperability by mapping one pair of elements at a time.

Keywords: application profiles; Dublin Core; Dublin Core Terms; semantic interoperability; MARC; metadata crosswalks

1. Application Profiles and Metadata Mapping

A preservation society in Ohio has just digitized some old photographs of Chillicothe, the state capital from 1803 until 1810 and the home of Majestic Theater, which has operated continuously for over a century and a half and has hosted many famous vaudeville performers, including Laurel and Hardy and Milton Berle. To make these images accessible to students and local history buffs, volunteers create a Dublin Core description that includes a title, description, and subject for each image, which renders them visible to automated harvesting utilities. But since this is a curated set of images about a particular place, the description could be enhanced with a record that describes the entire collection, using vocabulary from the Dublin Core Collection (DCMI, 2007) application profile, which includes a statement about access rights, pointers to associated collections, and a description of how the collection is accrued.

An *application profile* is a "declaration of the metadata terms an organization, information resource, application, or user commuity uses in its metadata," according to Greenberg and Severiens (2007), and is motivated by the need to enhance the discovery of a resource by diverse groups of people. In our hypothetical but realistic example, the owners of the images want to make their resources accessible to students or the curious public in a way that also preserves a piece of the historical record for future scholars. At the 2007 International Conference on Dublin Core and Metadata Applcations, project leaders from four continents reported on the design and use of application profiles to serve similar needs. For example, the SCROL (Singapore Cultural Resources Online) project designed a profile for managing access to images from multiple databases controlled by museums and archives (Wu, et al, 2007). And the DRIADE project (Digital Repository of Information and Data for Evolution) developed a profile for the management of heterogeneous data relevant to the study of evolutionary biology (Carrier, et al, 2007).

Though these projects have achieved varying degrees of technical maturity, most acknowledge the seminal work of Heery and Patel (2000), who characterize the application profile as a formalism that resolves the conflict between two groups of stakeholders. On the one hand, standards developers want to encourage consistency and continuity; on the other, application developers require flexibility and responsiveness. To meet the needs of both groups, Heery and Patel describe guidelines for the creation of application profiles, which may:

- *Draw on one or more existing namespaces.* Technically, a namespace is an element defined in an XML schema, though it is often understood to refer to a named domain containing a list of terms that could be, but is not yet, expressed in a formal syntax. In our scenario, elements such as *title* or *description* belong to the Dublin Core namespace, while elements such as *accrual method* belong to the Dublin Core Collections namespace. Additional namespaces can be added if they are required for a more detailed description. For example, if the digitized photos are used in a high-school course on the history of Ohio, the description might be enhanced with an element such as *audience* from the Gateway to Educational Materials (GEM, 2008) namespace, whose value would specify that this resource is appropriate for high-school juniors and seniors.
- *Refine standard definitions—but only by making them narrower, not broader*. For example, the GEM *audience* element is intended to annotate the grade level of a resource that can be used in a classroom. But since *audience* is a specialized description, it is formally linked to *description,* an element defined in the Dublin Core namespace that can replace it when a less detailed record is required. Because of this restriction on how definitions can be refined, the application profile permits complementary operations on the elements that comprise it. An element is *refined* or *replaced* when the element with the narrower meaning substitutes for the corresponding broader one. And an element is *dumbed down* when the element with the broader meaning is used instead.
- *Introduce no new data elements.* Data elements may not be added to existing namespaces, but may only be introduced into a description by including more namespaces, as we've indicated. To extend our example, suppose the historical society needed to keep track of where the records describing the digitized images reside in a local database. If so, a metadata standards expert could define a namespace such as *ChillicotheHistoricalSociety*, which might contain a *database-id* element, and add it to the application profile.

The technical infrastructure of the application profile addresses the needs of standards makers by creating incentives to use existing descriptive frameworks instead of creating new ones, preserving some degree of interoperability among records that describe similar resources. For systems designers, the application profile permits complex descriptions to be built up or collapsed using easily formalized operations.

In this paper, we show how the machinery of the application profile defined by Heery and Patel aids in the efficient management of metadata formats that are translated to and from MARC in OCLC's Crosswalk Web service (Godby, Smith and Childress, 2008), a utility that powers the metadata translation functions in OCLC Connexion® Client and a growing number of other products and services. The focus of our effort is the relationship between MARC and Dublin Core Terms (hereafter, DC-Terms) (DCMI, 2008), a namespace and de-facto application profile that extends Unqualified Dublin Core (hereafter DC-Simple) by adding elements such as *AudienceLevel* or *Mediator* and by refining DC-Simple elements such as <dc:relation> with *isReferencedBy* or *isReplacedBy*.

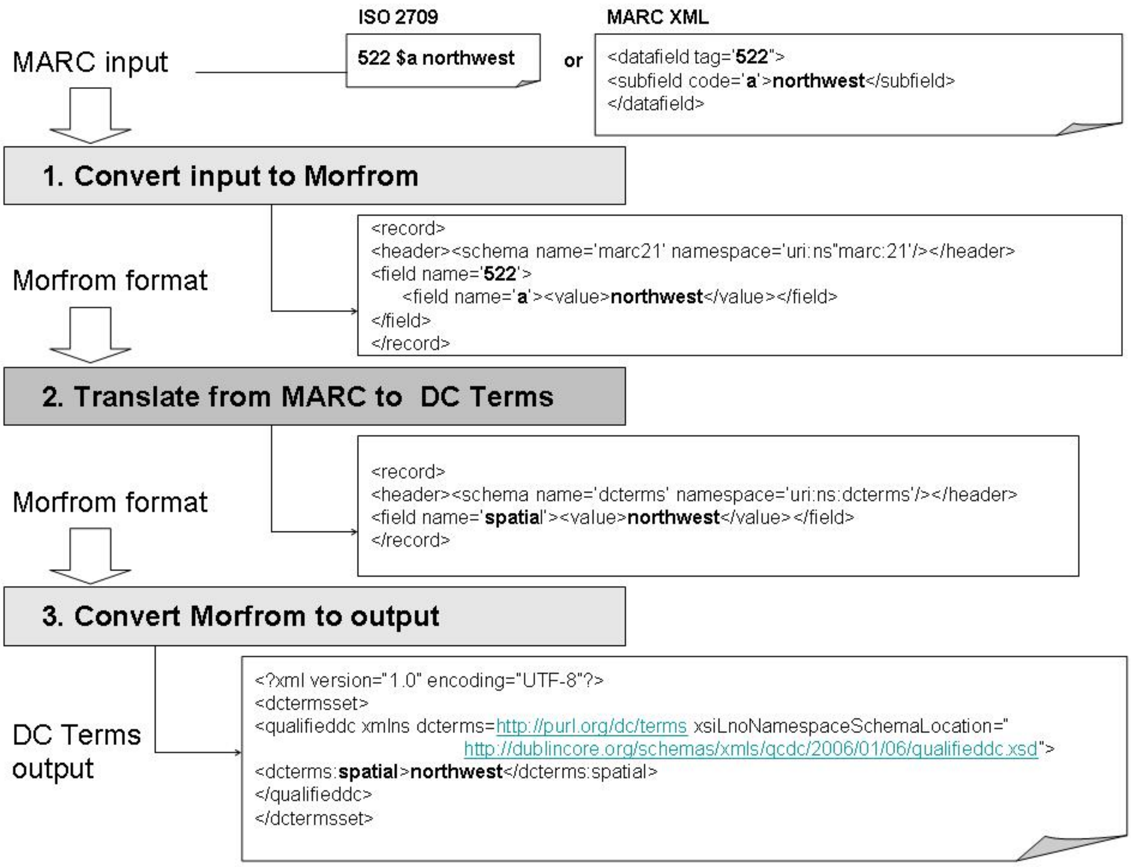
To implement the relationship between MARC and DC-Terms, we need to solve three problems. First, since the only publicly accessible crosswalk (LOC, 2008) was last updated in 2001 and has been defined only for one direction, from MARC to DC-Terms, we need to expedite the process of acquiring translation logic from metadata standards experts and converting it to executable code. Second, we need to manage versions. Software that exploits the inheritance structure in an application profile can process records conforming to DC-Terms and DC-Simple schemas, or to these schemas with local extensions. Given that MARC can be extended in similar fashion, and that input or output records may have multiple structural realizations—as XML, ISO-2709, or RDF, among others—the number of record variants to be translated can quickly explode. Finally, we need to manage change because standards are always evolving, as are the use cases that invoke them.

The solutions to all three problems emerge from the fact that the application profile, as well as the translation model underlying the Crosswalk Web Service, focus on the goal of achieving element-level interoperability, as described in the recent surveys by Chan and Zeng (2006) and Zeng and Chan (2006). The resulting is strikingly simple. The metadata subject matter expert edits a spreadsheet, from which the corresponding executable code is automatically generated. As a consequence, about two dozen types of records can be processed in a software environment that is rarely touched by human hands, eliminating a software maintenance problem in the MARC-to-Dublin Core crosswalk with a model that can eventually be applied to other relationships.

2. The DC-Terms application profile in the Crosswalk Web Service

Figure 1 illustrates the process flow for the translation of a record by the Crosswalk Web Service. As shown at the top of the figure, the input is a small (and invalid) MARC record consisting of a single field and subfield, *522 a Northwest*, encoded either in the MARC XML (LOC, 2007b) or ISO-2709 (ISO, 2008) syntax. The output is an XML-encoded DC-Terms record containing the element *DCTerms:spatial*, shown at the bottom of the figure. In Step 1, a utility program that we call a *reader* converts the native MARC input to a standardized, easy-to-process XML container syntax that we call Morfrom. Step 2 translates this record to a Morfrom representation of DC-Terms. In Step 3, a utility called a *writer* converts this result to an output syntax (here, another XML encoding) that has been formally defined by the Dublin Core standards community for DC-Terms.

These are the major operations of the core business logic in the Crosswalk Web Service. For a more technical discussion of the processing and data models, as well as the arguments that motivate this design, the reader is referred to our most recent article (Godby, Smith and Childress, 2008).


ISO 2709
MARC XML
MARC input
or
1. Convert input to Morfrom
Morfrom format
2. Translate from MARC to DC Terms
Morfrom format
3. Convert Morfrom to output
DC Terms output

```
522 $a northwest
```

```
<datafield tag='522">
<subfield code='a'>northwest</subfield>
</datafield>
```

```
<record>
<header><schema name='marc21' namespace='uri:ns"marc:21'/></header>
<field name='522'>
    <field name='a'><value>northwest</value></field>
</field>
</record>
```

```
<record>
<header><schema name='dcterms' namespace='uri:ns:dcterms'/></header>
<field name='spatial'><value>northwest</value></field>
</record>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<dctermsset>
<qualifieddc xmlns dcterms=http://purl.org/dc/terms xsiLnoNamespaceSchemaLocation="
                    http://dublincore.org/schemas/xmls/qcdc/2006/01/06/qualifieddc.xsd">
<dcterms:spatial>northwest</dcterms:spatial>
</qualifieddc>
</dctermsset>
```

FIG. 1. Data flow in the Crosswalk Web Service.

Since our present focus is on application profiles, the most important step in the model is Step 2, where the translation from MARC to Dublin Core Terms takes place. Three critical issues must be addressed. How does the metadata subject matter expert communicate the translation logic to the implementation? How is the translation implemented? And how does the application profile interact with the translation?

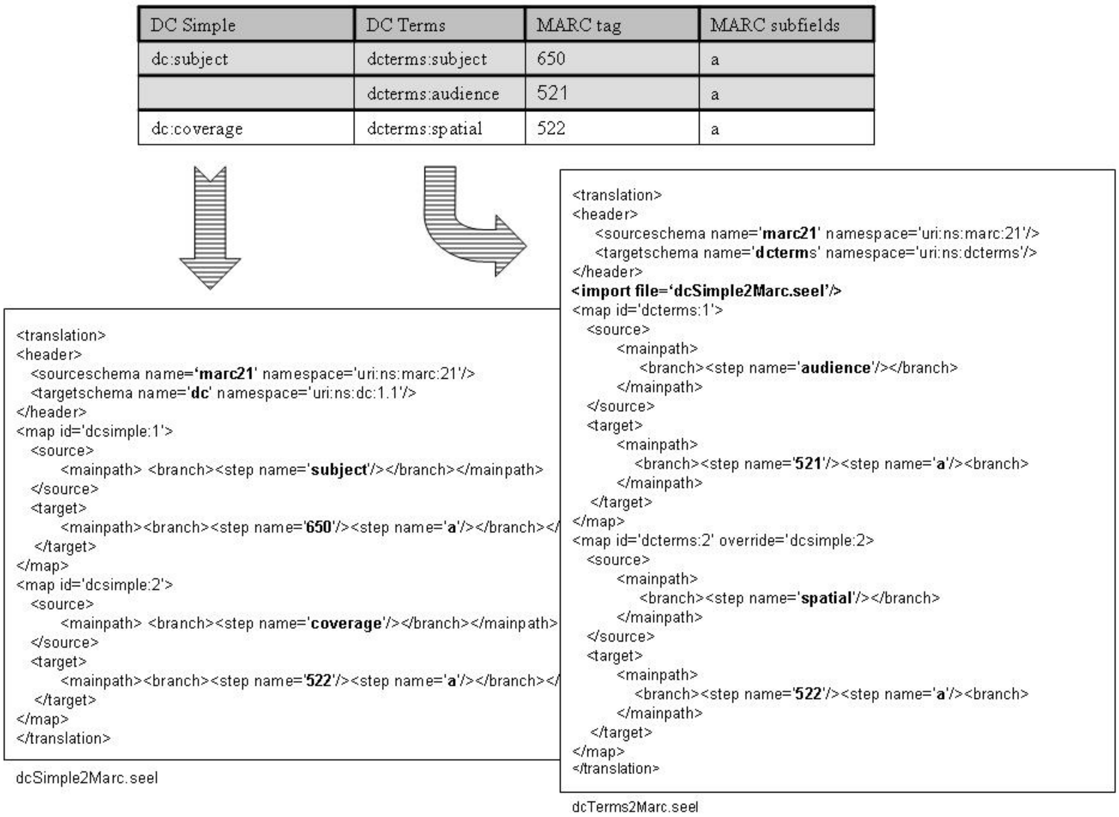
To answer the first question, the metadata standards expert fills out a spreadsheet like the one shown at the Figure 3, which shows the most important elements for implementing a crosswalk involving an application profile when the corresponding elements are related by a straightforward lexical substitution. Such is the case for our sample record shown in Figure 1, where *522 a* in the input is replaced by *spatial* in the output, a relationship that is expressed in the last row of the table in Figure 3. (More subtle relationships can also be expressed, as we will discuss shortly.) In the same row, the standards expert has recorded two additional facts about the Dublin Core-to-MARC relationship: that *dc:coverage* also maps to MARC 522 a, and that *dcterms:spatial* is 'dumbed down' to *dc:coverage* when a DC-Terms record requires a Dublin Core Simple manifestation. Thus, if the user with the record shown in Figure 1 had specified an output of DC-Simple instead of DC-Terms, the result would have been the record shown in Figure 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<simpledc xmlns dc='http://purl.org/dc/elements/1.1/>
<dc:coverage>northwest</dc:coverage>
</simpledc>
```

FIG. 2. A DC-Simple record.

The other rows in Figure 3 contain less complex information and are greyed out because they mention elements that are not represented in our sample record. In the first two rows, the names for the DC-Simple and DC-Terms are the same, so any effect of the dumb-down operation would be invisible. In the third row, *dcterms:audien*ce is not mapped to any corresponding DC-Simple element because it represents an element definition in the DC-Terms namespace that extends the descriptive scope of DC-Simple.

After the standards expert has filled out or edited the crosswalk spreadsheet, a software developer runs a Perl script against it to produce executable code. A sample is shown in the two boxes at the bottom of Figure 3. This is Seel (or Semantic Equivalence Expression Language) code, which we have designed, along with a program that interprets it, to model the information found in a typical crosswalk and make it actionable. A Seel script, expressed in XML, corresponds to a translation. The most important elements are <map>, which is a self-contained representation of a single row in a crosswalk; <source>, which identifies the input element, in this example, *522 a*; and <target>, which identifies an output element such as *coverage* or *spatial*. The <mainpath> element defines a path in the Morfrom record where the data of interest is located. In our sample record, the data *northwest* is found in a path ending with the elements *522 a* and will be written to a path containing elements named *coverage* or *spatial*. Finally, in addition to a set of maps, a Seel script must have a <header>, which lists locally defined URIs where technical specifications for the source and target schemas can be resolved.



DC Simple	DC Terms	MARC tag	MARC subfields
dc:subject	dcterms:subject	650	a
	dcterms:audience	521	a
dc:coverage	dcterms:spatial	522	a

```
<translation>
<header>
  <sourceschema name='marc21' namespace='uri:ns:marc:21'/>
  <targetschema name='dc' namespace='uri:ns:dc:1.1'/>
</header>
<map id='dcsimple:1'>
  <source>
       <mainpath> <branch><step name='subject'/></branch></mainpath>
  </source>
  <target>
       <mainpath><branch><step name='650'/><step name='a'/></branch></
   </target>
</map>
<map id='dcsimple:2'>
  <source>
       <mainpath> <branch><step name='coverage'/></branch></mainpath>
  </source>
  <target>
       <mainpath><branch><step name='522'/><step name='a'/></branch></
   </target>
</map>
</translation>
```

dcSimple2Marc.seel

```
<translation>
<header>
   <sourceschema name='marc21' namespace='uri:ns:marc:21'/>
   <targetschema name='dcterms' namespace='uri:ns:dcterms'/>
</header>
<import file='dcSimple2Marc.seel'/>
<map id='dcterms:1'>
  <source>
       <mainpath>
          <branch><step name='audience'/></branch>
       </mainpath>
  </source>
  <target>
       <mainpath>
         <branch><step name='521'/><step name='a'/><branch>
       </mainpath>
   </target>
</map>
<map id='dcterms:2' override='dcsimple:2>
  <source>
       <mainpath>
          <branch><step name='spatial'/></branch>
       </mainpath>
  </source>
  <target>
       <mainpath>
         <branch><step name='522'/><step name='a'/><branch>
       </mainpath>
   </target>
</map>
</translation>
```

dcTerms2Marc.seel

FIG. 3. A spreadsheet format and corresponding Seel maps.

Both Seel scripts shown in Figure 3 are automatically generated when the Perl script is applied to the spreadsheet. Both Seel files are complete scripts in the sense that they can be executed 'as is,' but a useful translation would usually contain many more maps than than those shown here. The script on the left is called whenever a client requests DC-Simple output from the Crosswalk Web Service. The script on the right is called when users ask for DC-Terms records.

Before we proceed with the discussion, it is instructive to highlight the most important features of these translations. The first script contains the DC element *subject*, which does not appear in the second. The second script contains a map to the DC-Terms element *audience*, which does not appear in the first. Both scripts contain maps to the DC elements from MARC 522 that represent geopatial information—*coverage* and *spatial*—at different levels of granularity. These maps and their relationships to each other in the two translations are sufficient to illustrate the essential operations that must be modeled in the execution of an application profile in a translation. How is an element dumbed down? How is an element added to a namespace? How is the meaning of an element in one namespace overridden by an element in another? And how is an instance record conforming to the application profile, which must, by definition, contain a mixture of elements from different namespaces, produced in a translation?

Conceptually, the dumb-down operation is the easiest. To map *coverage* from MARC 522 a instead of *spatial*, the client needs only to run the dcSimple2MARC script (on the left in Figure 3) instead of dcTerms2MARC. To dumb down an original Dublin Core Terms record instead of a MARC record, the user submits DC-Terms input to the Crosswalk Web Service and specifies DC-Simple as the output. Internally, the Web service would translate DC-Terms to MARC and then MARC to DC-Simple using the corresponding translations from the other direction that have been generated in the same manner as those we discuss here.

The other operations are implemented through elements defined in the Seel language. To describe how maps to elements in multiple namespaces are added to the translation and how the translated record ends up with a mixture of elements from different sources, we need to point out a detail that appears in the DC-Terms script but is not present in the DC-Simple script. The <import> element, shown in bold type in Figure 3, forces the inclusion of the maps defined in the DC-Simple script to create a comprehensive translation that consists of maps from both scripts. One effect of the *impor*t operation is to extend the maps involving the base standard—here, DC Simple—with a set of elements from a different namespace. In our example, the map to *dcterms:audience* appears in the DC-Terms output as the only new extension. But the spreadsheet and corresponding Seel script that represent the full application profile contains many more elements coded in this pattern.

Another effect of the <import> element on the DC-Terms script is to propagate into the translated record the binding of the *audience* element to the DC-Terms namespace and that of the *subject* element to the DC-Simple namespace. Recall that the <header> element in the DC-Terms script specifies *dcterms* as the namespace of the target, ensuring that the *dcterms* namespace is attached to the topmost element and is inherited by the target elements of each map in the translation—here, *spatial* and *audience*. An analogous operation happens when the DC-Simple script is executed. But when dcSimple2Marc.seel is imported into the DC-Terms script, the DC-Simple namespace is explicitly attached to every element that appears in DC-Simple but not in DC-Terms by a local *namespace* attribute, which overrides the default *dcterm*s namespace that would have otherwise been assigned. Below is a DC-Terms record that contains the elements *audience* and *subject*. The Morfrom representation of a record produced by the application of the Seel scripts in Figure 3 is shown first (with the explicit *namespace* element shown in bold), then the DC-Terms XML syntax produced by one of the writer utilities in the Crosswalk Web Service. Note the treatment of the <subject> element, shown in bold at the bottom of Figure 4, which demonstrates that the Morfrom element with a namespace attribute is represented in the output as an element from from the DC-Simple namespace, not the DC-Terms default.

```
<record>
<header><schema name=”dcterms” namespace=’uri:ns:dcterms’/></header>
<field name=’audience’>high school students</field>
<field name=’subject’ namespace=’uri:ns:dc:1.1’><value>geography</value></field>
</record>
```

```
<?xml version=“1.0” encoding=“UTF-8”?>
<dctermsset>
<qualifieddc xmlns dcterms=”http://purl.org/dc/terms”, xmlns dc=http://purl.org.dc/elements/1.1/
xsi:noNamespaceSchemaLocation=“
                              http://dublincore.org/schemas/xmls/qcdc/2006/01/06/qualifieddc.xsd”>
<dc:subject>geography</dc:subject>
<dcterms:audience>high school students</dcterms:audience>
</qualifieddc>
</dctermsset>
```

FIG. 4. Morfrom and native XML encodings of a DC-Terms record.

The Seel scripts in Figure 3 implement the smallest possible application profile, which contains elements from two namespaces, but this model can be extended if necessary. The DC-MARC translation at OCLC requires an additional set of elements that keep track of locally defined identifiers and other houskeeping information. To manage them, we defined elements in the *OCLC-Admin* namespace, mapped them to DC-Terms using a modified version of the spreadsheet like that shown in Figure 3, and created the corresponding Seel scripts. Now the same spreadsheet can serve as input to four translations, all easily maintained because they are automatically generated: MARC to DC-Simple, MARC to DC-Terms, MARC to OCLC-Simple, which uses the DC-Simple and OCLC-Admin namespaces; and MARC to OCLC-Terms, which uses all three namespaces.

The last operation is the override, the inverse of dumb-down. In our example, *spatial* overrides *coverage* in a DC-Terms record because it offers the chance for a more precise definition of a piece of geospatial data. The critical code is in the first map of the DCTerms2MARC script, shown on the right in Figure 3. Because <map> elements are self-contained and modular, they can carry *id* attributes that uniquely index them. Though this string can be any value, the maps in Figure 3 have straightforward names: dc-simple:1, dc-simple:2, dc-terms-1 and dc-terms-2. The map with the id value of *dc-terms:1*, which translates the *spatial* element, also has an *override* value of *dc-simple:1*, thus associating it with the map whose target is *coverage*. Now that the dominance relationship between the two maps is established, the DC-Terms map involving *spatial* is executed instead of the DC-Simple map containing *coverage* when dcterms2MARC.seel is executed.

The *dcterms:audience* element is another candidate for the override operation. When it is implemented as shown in Figure 3, it is interpreted as one of the DC-Terms elements that extends the description of DC-Simple, as we have discussed. But some members of the Dublin Core community have proposed that *dcterms:audience* should override *dc:description*. To implement this change, the metadata standards expert would need only fill in the blank box in the second row of the spreadsheet, and the *override* attribute with the appropriate value would be automatically added to the map whose id value is *dc-terms:2*, the second map in the dcTerms2MARC translation.

3. A more realistic example

Before leaving the technical discussion, we need to point out that the spreadsheet shown in Figure 3 is much too simple for realistic data. When the relationships required for managing application profiles are stripped out, the spreadsheet does little more than model a one-to-one map of source to target elements: *521 a* to *audience*, *650 a* to *subject*, and so on. More complex relationships are usually required for mapping the bibliographic metadata that pass through

OCLC’s systems. For example, the mapping between source and target may be conditional on the value of data or the presence or absence of particular fields elsewhere in the record. And sometimes the data itself must be manipulated or identified with a special encoding scheme. Figure 3 shows the rest of the spreadsheet that the metadata standards expert fills out to create a production-quality MARC-to-Dublin Core translation. The greyed-out columns are the same as those in Figure 3 and indicate that the MARC tag 050 maps to *subject* regardless of whether it is interpreted as a member of the DC-Terms or DC-Simple namespace. The three columns on the right contain prompts that instruct the automated process to join the elements in MARC subfields *a* and *b* with a space and to execute this translation only if both indicators are present. The column labeled *XSI Type* specifies that the output data be interpreted as a LCC number, an encoding scheme listed in the DC-Terms namespace.

DC Simple	DC Terms	XSI Type	MARC tag	Indicators	Subfields	Special rule
Subject	Subject	dcterms:LCC	050	??	a,b	join( )

FIG. 5. An extended spreadsheet.

The Seel map created from this entry is shown in Figure 6. The additional details, shown in bold type, include the <context> element that checks for the existence of the indicators; a <value> element on the <source> that joins the subfields; and a <value> element on the <target>, whose attribute identifies the data as an LCC encoding. This example shows that even when a Seel map is expressive enough to model real-world relationships, the code remains fairly legible. The logic can be still represented transparently in a spreadsheet that is maintained by a non-programmer and automatically converted to executable code.

```
<translation>
   <header>
      <sourceschema name=’marc’ namespace=’uri:ns:marc:21’/>
      <targetschema name=’dc’ namespace=’uri:ns:dc:1.1’/>
   </header>
   <map id=3’>
      <source>
        <mainpath>
           <branch bid=’1’><step name=’050’>
              <value><join with=’  ’ include=’a,b’></join></value>
            </branch>
        </mainpath>
        <context bid=’1’>.
          <exists><path><step name=”i1”/></exists>
          <exists><path><step name=”i2”/></exists>
        </context>
     </source>
     <target>
       <mainpath>
          <branch bid=’1’>
                <step name=’subject’><value type=’http://purl.org/dc/terms/LCC/’/></step>
          </branch>
      </mainpath>
     </target>
   </map>
</translation>
```

FIG. 6. The Seel map generated from the extended spreadsheet.

When this script is applied to a MARC record containing the fragment ‘050 ## $a PS3537.A618 $b A88 1993,’ it produces the Morfrom and Dublin Core output shown in Figure 7. Note that the final outcome is a DC-Simple record, but the *xsi:type* attribute on the <subject> element properly identifies the encoding scheme of the data from the DC-Terms namespace.

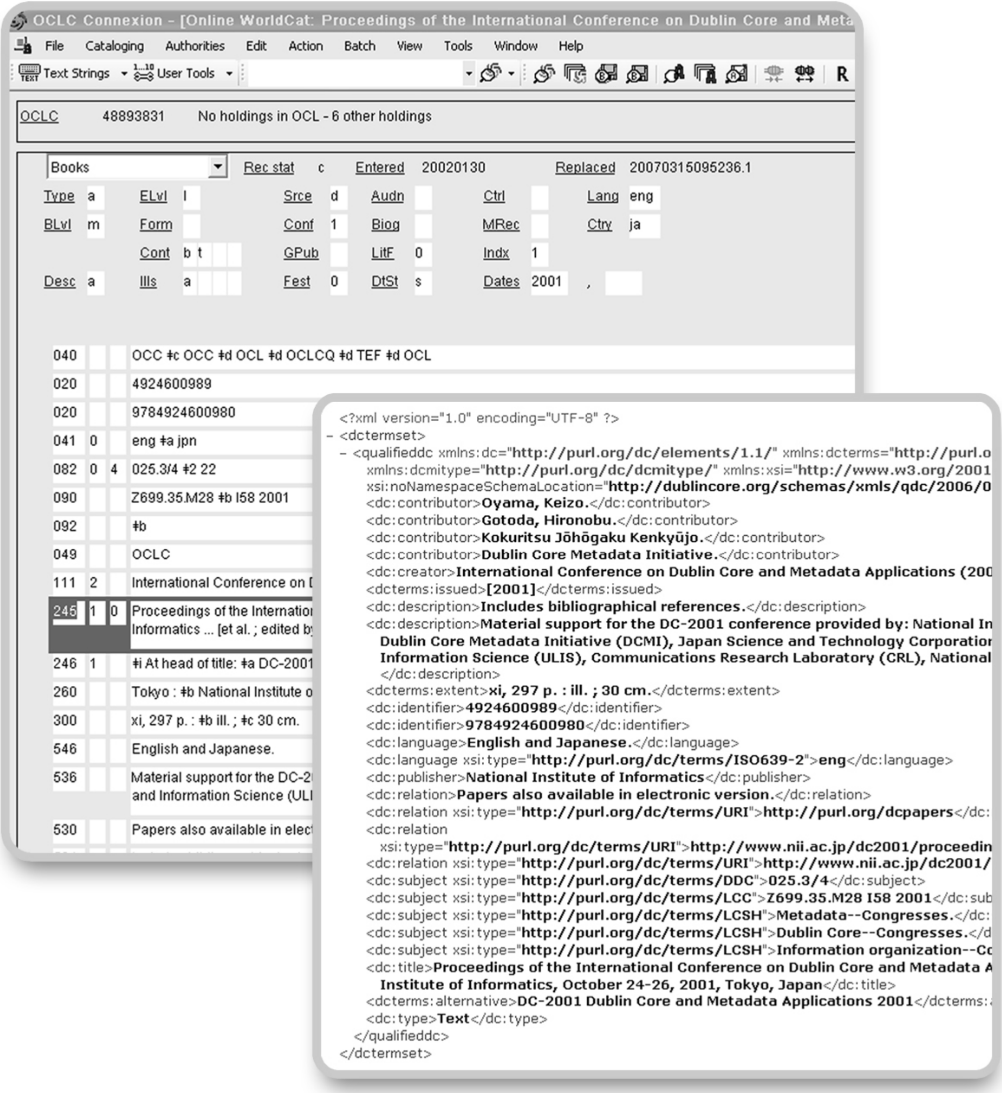
```
<?xml version="1.0"?>
<record>
<header><schema name="dc" namespace="uri:ns:dc:1.1"/></header>
<field name="subject">
        <value type="http://purl.org/dc/terms/LCC">PS3537.A618 A88 1993</value>
</field>
</record>
<?xml version="1.0" encoding="UTF-8"?>
```

```
<simpledc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"
xsi:noNamespaceSchemaLocation="http://dublincore.org/schemas/xmls/qdc/2006/01/06/simpledc.xsd">
<dc:subject xsi:type='http://purl.org/dc/terms/LCC'>PS3537.A618 A88 1993</dc:subject>
</simpledc>
```

FIG.7. Morfrom and native XML encodings of a DC-Simple record with a refined element.

4. Summary and future work

To summarize, we have shown how the DC-Terms application profile is invoked in a non-trivial use case: the translation of bibliographic metadata in a high-volume production environment. Users can test the results by submitting their own records to the public demo on the OCLC ResearchWorks page (OCLC, 2008). Or they can use the Dublin Core *export* functions in the OCLC Connexion® Client, as shown in Figure 8.


OCLC Connexion - [Online WorldCat: Proceedings of the International Conference on Dublin Core and Meta
File Cataloging Authorities Edit Action Batch View Tools Window Help
Text Strings User Tools
R

```
OCLC      48893831      No holdings in OCL - 6 other holdings

Books        Rec stat  c    Entered  20020130        Replaced  20070315095236.1
Type  a    ELvl  I       Srce  d    Audn         Ctrl        Lang  eng
BLvl  m    Form          Conf  1    Biog         MRec        Ctry  ja
           Cont  b t     GPub       LitF  0      Indx  1
Desc  a    Ills  a       Fest  0    DtSt  s      Dates 2001  ,

040         OCC ‡c OCC ‡d OCL ‡d OCLCQ ‡d TEF ‡d OCL
020         4924600989
020         9784924600980
041  0      eng ‡a jpn
082  0  4   025.3/4 ‡2 22
090         Z699.35.M28 ‡b I58 2001
092         ‡b
049         OCLC
111  2      International Conference on [
245  1  0   Proceedings of the Internatio
            Informatics ... [et al. ; edited b
246  1      ‡i At head of title: ‡a DC-2001
260         Tokyo : ‡b National Institute o
300         xi, 297 p. : ‡b ill. ; ‡c 30 cm.
546         English and Japanese.
536         Material support for the DC-2
            and Information Science (ULI
530         Papers also available in elec
```

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <dctermset>
  - <qualifieddc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.o
      xmlns:dcmitype="http://purl.org/dc/dcmitype/" xmlns:xsi="http://www.w3.org/2001
      xsi:noNamespaceSchemaLocation="http://dublincore.org/schemas/xmls/qdc/2006/0
      <dc:contributor>Oyama, Keizo.</dc:contributor>
      <dc:contributor>Gotoda, Hironobu.</dc:contributor>
      <dc:contributor>Kokuritsu Jōhōgaku Kenkyūjo.</dc:contributor>
      <dc:contributor>Dublin Core Metadata Initiative.</dc:contributor>
      <dc:creator>International Conference on Dublin Core and Metadata Applications (200
      <dcterms:issued>[2001]</dcterms:issued>
      <dc:description>Includes bibliographical references.</dc:description>
      <dc:description>Material support for the DC-2001 conference provided by: National In
        Dublin Core Metadata Initiative (DCMI), Japan Science and Technology Corporatio
        Information Science (ULIS), Communications Research Laboratory (CRL), National
        </dc:description>
      <dcterms:extent>xi, 297 p. : ill. ; 30 cm.</dcterms:extent>
      <dc:identifier>4924600989</dc:identifier>
      <dc:identifier>9784924600980</dc:identifier>
      <dc:language>English and Japanese.</dc:language>
      <dc:language xsi:type="http://purl.org/dc/terms/ISO639-2">eng</dc:language>
      <dc:publisher>National Institute of Informatics</dc:publisher>
      <dc:relation>Papers also available in electronic version.</dc:relation>
      <dc:relation xsi:type="http://purl.org/dc/terms/URI">http://purl.org/dcpapers</dc:
      <dc:relation
        xsi:type="http://purl.org/dc/terms/URI">http://www.nii.ac.jp/dc2001/proceedin
      <dc:relation xsi:type="http://purl.org/dc/terms/URI">http://www.nii.ac.jp/dc2001/
      <dc:subject xsi:type="http://purl.org/dc/terms/DDC">025.3/4</dc:subject>
      <dc:subject xsi:type="http://purl.org/dc/terms/LCC">Z699.35.M28 I58 2001</dc:sub
      <dc:subject xsi:type="http://purl.org/dc/terms/LCSH">Metadata--Congresses.</dc:
      <dc:subject xsi:type="http://purl.org/dc/terms/LCSH">Dublin Core--Congresses.</d
      <dc:subject xsi:type="http://purl.org/dc/terms/LCSH">Information organization--Cc
      <dc:title>Proceedings of the International Conference on Dublin Core and Metadata A
        Institute of Informatics, October 24-26, 2001, Tokyo, Japan</dc:title>
      <dcterms:alternative>DC-2001 Dublin Core and Metadata Applications 2001</dcterms:
      <dc:type>Text</dc:type>
    </qualifieddc>
  </dctermset>
```

FIG. 8. A DC-Terms record crosswalked from MARC and exported to a file using OCLC Connexion Client®.

Though the result is clean, the interest is not merely theoretical because the application profile solves a significant practical problem. The complex relationships among MARC, Dublin Core, and related namespaces resolve to a mutable set of translations involving some elements that require special definitions, some that are defined in public standards, and some that are required only for local maintenance—exactly what the application profile was designed for, according to Heery and Patel (2000). With its focus on elements that can be mixed, matched, added, and redefined, the concepts that make up the application profile are a natural fit with the translation model of the Crosswalk Web Service, which attempts to achieve interoperability one element at a time by mapping those with similar meanings and manipulating their content when necessary. Before application profiles were defined and a translation model was developed that enforces transparency and reuse, the MARC-Dublin Core relationship at OCLC was managed with a large and brittle collection of pairwise translations—from DC-Terms to DC-Simple, from MARC to DC-Simple, MARC to DC-Terms, MARC to OCLC-Terms, and so on—a collection that multiplied quickly when structural variation was factored in.

Nevertheless, we consider this implementation to be the first step to a more generic solution. We eventually hope to develop a user interface that accepts input from a Web-accessible form and produces two outputs: the executable Seel scripts, and the corresponding crosswalk formatted as a table for human consumption that is more abstract than the spreadsheet that must be maintained by metadata subject matter experts who still must be coached by our development staff. At a deeper level, we plan to exploit more of the element-oriented architecture in our translation processing, especially the URIs that are attached to each element and can carry information about the namespace it belongs to, the path to the element expressed in a formal syntax, and notes about local conditions. This is a large subject, but well worth attention because it will link our work to valuable implementations of metadata registries (Heery and Wagner, 2003) and annotation profiles (Palmér, et al., 2007). But the fact that we have achieved userful intermediate results and can envision a migration path to a more generic solution is a testament to the far-reaching consequences of Heery and Patel's original vision.

References

DCMI. (2007). *Dublin Core Collections Application Profile.* Retrieved April 10, 2008 from http://dublincore.org/groups/collections/collection-application-profile/index.shtml.

Carrier, Sarah, Jed Dube, and Jane Greenberg. (2007). The DRIADE project: Phased application profile development in support of open science. *Proceedings of the International Conference on Dublin Core & Medata Applications, 2007* (pp. 35-42).

DCMI. (2008). *DCMI metadata Terms.* Retrieved April 10, 2008, from http://dublincore.org/documents/dcmi-terms/.

GEM. (2008). *Gateway to 21st Century Skills.* Retrieved April 10, 2008, from http://www.thegateway.org/.

Gill, Tony, Anne J. Gilliland, and Mary S. Woodley. (n.d). *Introduction to metadata. Pathways to digital information.* Online Edition, Version 2.1. Retrieved June 10, 2008, from http://www.getty.edu/research/conducting_research/ standards/intrometadata/glossary.html#C.

Chan, Lois M. and Marcia Lei Zeng. (2006). Metadata interoperability and standardization - A study of methodology, Part I. *D-Lib Magazine, 12*(6). Retrieved April 10, 2008, from http://www.dlib.org/dlib/june06/chan/06chan.html.

Godby, Carol J., Devon Smith, and Eric Childress. (2008). Toward element-level interoperability in bibliographic metadata. *Code4Lib Journal, 1*(2). Retrieved April 10, 2008, from http://journal.code4lib.org/articles/54.

Greenberg, Jane, Kristina Spurgin and Abe Crystal. (2007). Functionalities for automatic-metadata generation applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics, and Ontologies, 1*(1), 3-20.

Heery, Rachel and Manjula Patel. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne, 25.* Retrieved April 10, 2008, from http://www.ariadne.ac.uk/issue25/app-profiles/.

Heery, Rachel and Harry Wagner. (2002). A metadata registry for the Semantic Web. *D-Lib Magazine 8*(5). Retrieved June 10, 2008, from http://www.dlib.org/dlib/may02/wagner/05wagner.html.

ISO. (2008). *ISO: 2709:1996.* Retrieved April 10, 2008, from http://www.iso.org/iso/iso_catalogue/ catalogue_tc/catalogue_detail.htm?csnumber=7675

LOC (2007a). *MARC 21 specifications for record structure, character sets, and exchange media*. Retrieved April 10, 2008, from http://www.loc.gov/marc/specifications/specchartables.html.

LOC. (2007b). *MARC XML: MARC 21 Schema.* Retrieved April 10, 2008, from http://www.loc.gov/ standards/marcxml.

LOC. (2008). *MARC to Dublin Core crosswalk.* Retrieved April 10, 2008, from http://www.loc.gov/marc/marc2dc-2001.html.

OCLC. (2008). *ResearchWorks: Things to play with and think about.* Retrieved April 10, 2008, from http://www.oclc.org/research/researchworks/default.htm.

Palmér, Matthias, Fredrik Enokkson, Mikael Nilsson and Ambjörn Naeve. (2007). Annotation profiles: Configuring forms to edit RDF. *Proceedings of the International Conference on Dublin Core & Medata Applications, 2007* (pp. 10-21).

Wu, Steven, Barbara Reed and Paul Loke. (2007). SCROL application profile. *Proceedings of the International Conference on Dublin Core & Medata Applications, 2007* (pp. 22-29).

Zeng, Marcia Lei and Lois M. Chan. (2006). Metadata interoperability and standardization - A study of methodology, Part II. *D-Lib Magazine, 12*(6). Retrieved April 10, 2008, from http://www.dlib.org/dlib/june06/zeng/06zeng.html

Relating Folksonomies with Dublin Core

Maria Elisabete Catarino
University of Minho
Portugal / Capes-MEC-Brazil
ecatarino@dsi.uminho.pt

Ana Alice Baptista
University of Minho, Portugal
analice@dsi.uminho.pt

Abstract

Folksonomy is the result of describing Web resources with tags created by Web users. Although it has become a popular application for the description of resources, in general terms Folksonomies are not being conveniently integrated in metadata. However, if the appropriate metadata elements are identified, then further work may be conducted to automatically assign tags to these elements (RDF properties) and use them in Semantic Web applications. This article presents research carried out to continue the project Kinds of Tags, which intends to identify elements required for metadata originating from folksonomies and to propose an application profile for DC Social Tagging. The work provides information that may be used by software applications to assign tags to metadata elements and, therefore, means for tags to be conveniently gathered by metadata interoperability tools. Despite the unquestionably high value of DC and the significance of the already existing properties in DC Terms, the pilot study show revealed a significant number of tags for which no corresponding properties yet existed. A need for new properties, such as Action, Depth, Rate, and Utility was determined. Those potential new properties will have to be validated in a later stage by the DC Social Tagging Community.

Keywords: folksonomy; social tagging; metadata; Dublin Core

1. Dublin Core and Folksonomies: the Context

The highly active participation of users in the construction and organization of Internet contents arises from the evolution of the technologies used in the Web, the so-called Web 2.0. It is "the network as platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an 'architecture of participation', and going beyond the page metaphor of Web 1.0 to deliver rich user experiences". (O'Reilly, 2005).

Among the new possibilities of the Web 2.0 folksonomy comes up as "the result of personal free tagging of information and objects (anything with an URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information" (Wal, 2006). The tags which make up a folksonomy would be key-words, categories or metadata (Guy; Tonkin, 2006). In this brief definition of tag, it can be noticed that tags can play different roles.

Folksonomies describe the Web resources and as such it may be expectable that they are intelligible by machines and thus used by Semantic Web applications. To do so, properties (also known as "RDF links") are needed in order to clarify and express how given tags relate to the resource they describe. The DC Terms properties (from now on only referred as DC properties) are of high value to be used as a basis for interoperability and their wide acceptability is a good measure of this value. However, they are oriented to describing resources from the classical standpoints of authors and libraries, whereas in Web 2.0, resources are described from the highly diverse perspective of users.

The project Kinds of Tags (KoT) focuses its attention "on the analysis of tags that are in common use in the practice of social tagging, with the aim of discovering how easily tags can be 'normalised' for interoperability with standard metadata environments such as the DC Metadata Terms" (Baptista et al., 2007). Within KoT it was observed that there are some tags to which none of the existing DC properties could be adequately assigned. This indicates that other metadata elements might need to be identified. Preliminary results from this project were presented in DC-2007 and NKOS-2007 describing some probable new elements: Action_Towards_Resource, To_Be_Used_In, Rate and Depth (Baptista et al., 2007 and Tonkin et al., 2007).

In order to continue this analysis a deeper and more detailed research in underway and it aims to answer the following questions:

- do the DC properties have the necessary semantics to clarify and express how given tags relate to the resource they describe?
- if not, which other properties that hold this semantics can be identified to complement DC and to be used in social tagging applications?

This research uses the same data set that was used in KoT and begun with a detailed pilot study regarding the tags of the first five resources of the data set. This article presents the results of the pilot study and also refers some preliminary results of the final study. These indicate that some new properties may be needed for social tagging applications, which implies the possible construction of an application profile to be proposed to the Social Tagging community.

2. The Research Project: an in-depth Study following up KoT Preliminary Results

The dataset used in this project is the same of KoT: it is composed of 50 records of resources which were tagged in two systems of social bookmarking: Connotea and Delicious. Each record is composed has information distributed in two groups of data: a) data related to the resource as a whole: URL, number of users, research date; and b) data related to the tags assigned to the resource: social bookmarking system, user's nickname, bookmarked date and the tags.

A relational database was set up with the DCMI Metadata Terms and the KoT data set that was imported from its original files. The following tables were created: Tags, Users, Documents, Key-tags and Metadata.

There is a total number of 5098 tags (Connotea: 901; Delicious: 4819). The total number of users amounts to 15.381 (Connotea: 509; Delicious: 14.872). Considering that different users in different resources repeatedly assigned a tag, there is a total of 75.429 tag occurrences (Connotea: 3.698; Delicious: 71.731). It is important to consider the total number of tag occurrences, since a tag could correspond to different metadata elements depending on the resource to which it was assigned (for instance, a tag could correspond to Subject in a resource and to Title or Description in another).

The whole study is made manually in order to be as precise as possible regarding the meaning of the tags. It was divided in four stages: 1 – Analysis of tags; 2 – Identification of complementary properties; 3 – Formalization of the new properties in an ontology-like representation; 4 – Validation by the community and release of the first version of the proposal.

The first stage consists of an analysis of all tags contained in the dataset. At this stage all tags assigned to the resources are analysed, grouped in what we call key-tags and then DC properties are assigned to them when possible. A Key-tag is a normalised tag that represents a group of similar tags. For instance, the key-tag Library Science stands for tags `library.science`, `library_science` or `library-science`.

Once that the meaning of tags is not always clear, it is necessary to dispel doubts by complementarily turning to lexical resources (dictionaries, encyclopedias, Word Net, Wikipedia, etc), and analyzing other tags of the same users. Contacting the users may be a last alternative to try to find out the meaning of a given tag.

The second stage aims at proposing complementary properties to the ones already existing in the DCMI Metadata Terms (DCMI Usage Board, 2008). Key-tags to which none DC property was assigned in stage one will now be subject to further analysis in order to identify new properties specific to Social Tagging applications. This analysis takes into account all DC standards and recommendations, including the DCAM model, the ISO Standard 15836-2003 and the NISO Standard Z39.85-2007.

The next stage comprises the adaptation of an already existing DC ontology-like representation of the DC elements and their semantics. This will make use of Protégé, an ontology editor developed at Stanford University. The ontology will be encoded in OWL, a language endorsed by the W3C.

Finally, the fourth stage intends to submit a proposal for a DC Social Tagging application profile to the DC social tagging community for comments and feedback via online questionnaires. After this phase, a first final version of the proposal will be submitted to the community.

A pilot study was conducted for the first two stages with the first five resources of the data set. It allowed us to refine the proposed methodology and, in the first stage, to verify whether the proposed variants for grouping and analyzing tags are adequate. In the second stage, the pilot study allowed to have a preliminary overview of the percentage of tags to which DC properties could be assigned and, complementarily, the percentage of tags that would fit in new properties. As it was impossible to determine the meaning of some tags, there is a high percentage of non-assigned tags.

An important concern regarding tag analysis is the fact that as tags are assigned by the resources' users, that inevitably leads to a lack of homogeneity in their form. Therefore, it was necessary to establish some rules in order to properly analyze tags, establish key-tags and relate DC properties with them.

3. Rules for the first two stages

3.1. Rules for the first Stage

The first rule to be observed concerns the alphabet. In this project, only tags written in Latin alphabet were considered. Further studies should involve the analysis of tags written in different alphabets. For example: Greek/Ελληνική, cyrillic/Кирилица, Chinese/中國, Japanese/日本語, etc.

Another rule is related to language. The dataset comprises tags written in different languages. It was possible to identify and translate 425 tags written in languages other than English, which corresponded to 8,3% of the total number of tags as shown in TABLE 1.

TABLE 1. Number of identified and translated tags in languages other than English.

ISO 639 acronym	**Language**	**No. of tags**	**ISO 639 acronym**	**Language**	**No. of tags**
CA	Catalan	43	**HU**	Hungarian	9
CS	Czech	3	**IT**	Italian	16
DA	Danish	3	**MUL**	Multiple Languages[1]	57

[1] Tags that have the same spelling in several languages.

ISO 639 acronym	Language	No. of tags	ISO 639 acronym	Language	No. of tags
DE	German	51	**NL**	Dutch	16
ES	Spanish	47	**NO**	Norwegian	9
ET	Estonian	2	**PL**	Polish	2
EU	Basque	1	**PT**	Portuguese	77
FI	Finnish	9	**RO**	Romanian	4
FR	French	68	**SV**	Swedish	8
HR	Croatian	1	**TR**	Turkish	1

Most of the tags were, however, written in English. Thus, English was the chosen language to represent Key-tags.

Depending on the Key-tags, certain criteria concerning the classification of words need to be established: simple or compound, singular or plural, based on a thesaurus structure in its syntactical relations. In these cases, the rules to establish thesauri structure were followed as indicated by ISO 2788-1986 Standard.

It was still necessary to create rules to deal with compound tags, as they contain more than one word. There are two kinds of compound tags: (1) the ones that are related to only one concept and therefore originate only one key-tag (e.g. `Institutional Repositories`); and (2) the ones that are related to two or more concepts and therefore originate two or more key-tags (e.g. `digital-libraries:dublincore`).

In the first kind, compound tags are composed by a focus (or head) and a modifier (International Standards Organization, 1986). The focus, i.e. the noun component which identifies the general class of concepts to which the term as a whole refers, and the modifier, i.e. one or more components which serve to specify the extension of the focus; in the example above: `Institutional` (modifier) `Repositories` (focus). It is a compound term that comprises a main component or focus and a modifier that specifies it.

In the second kind, compound tags are related to two or more distinct Key-tags, as for example: `digital-libraries:dublincore`, which would be part of the group of two distinct Key-tags: `Digital Libraries` and `Dublin Core`. In this second segment there is not a relation of focus/difference between the components as they are totally independent.

3.2. Rules for the second stage

In the occurrence of Simple tags there is a peculiarity to be noticed that relates to the way tags are inserted in the social bookmarking sites: the way tags are inserted can interfere with the system's indexation. In Delicious the only separator is the space character and everything that is typed separated by spaces will be considered distinct tags. For example, if the compound term `Social Tagging` is inserted containing only the space as separator, the system will consider two tags: `Social` and `Tagging`. In order to be inserted as a compound tag it is necessary to use special characters such as underscore, dashes and colons. Some examples of such kind of compound tags are: `social+tagging`, `social_tagging`, `social-tagging`.

In Connotea tags are also separated by a space or a comma. However, Connotea suggests to users to type compound tags between inverted commas. For example, if the user inserts `Controlled Vocabularies` without placing the words between inverted commas, the words will be considered two distinct tags; however, if they are typed between inverted commas ("`Controlled Vocabularies`") the system will generate only one compound tag. This simple, yet important issue, has a high implication on the system's indexation of the tags.

To exemplify what is said above there is an example of a Delicious user who, when assigning tags to the resource "The Semantic Web", written by Tim Berners-Lee, inserted the following tags: `the`, `semantic`, `web`, `article`, `by`, `tim`, `berners-lee`, without using the characters of word combination (_ ; - etc). The system generated seven simple tags. However, it is clear that these tags can be post-coordinated[2] to have a meaning such as Title, Creator and Subject.

Thus, as a first rule, in the cases when simple tags could clearly be post-coordinated, they were analyzed as a compound term for the assignment of the DC Property. However, this analysis could only be carried out in relation to only one resource's user at a time and never to a group, since it can mischaracterize the assignment of properties.

The second rule concerns tags that correspond to more than one DC Property. It is considered two different situations: simple and compound tags. The easiest case is the one of simple tags. If simple tags occur to which two or more properties can be assigned, then all the properties are assigned to the tag. For example in the resource entitled "An Architecture for Information", the properties "Title" and "Subject" are assigned to the Key-tag `Architecture`.

As explained earlier, compound tags, however, can correspond to two or more key-tags. Thus the relationship with DC properties is made through the key-tags. These are treated as simple tags in the way they are related to DC properties. For example the tag `doi:10.1045/april2002-weibel`, corresponds to three Key-tags, `doi:10.1045`, `april 2002` and `Stuart L. Weibel`, each one of them corresponding to a different property: Identifier, Date and Creator (respectively). There may also be cases of compound tags that represent two different values for the same property, as in `folksonomiestagging`, that was splitted into two Key-tags: `Folksonomy` and `Tagging`, to which both the subject property was assigned.

Another rule is related to tags whose value corresponds to the property Title. Tags will be related to the element "Title" when they are composed by terms found in the main title of the resource. For example, `Folksonomies`, `WEb2.0`. Another example is the case of the resource entitled "`Social Bookmarking Tools`", where the tags `Social`, `Bookmarking`, `Tools`, that were assigned by the same user, and thus, are post-coordinated.

4. Tag Analysis

As stated earlier, this stage consists of an analysis of all tags contained in the dataset. At this stage all tags assigned to the resources are analyzed, grouped in key-tags and then DC properties are assigned to them when possible. In this stage it was necessary to use lexical resources (dictionaries, WordNet, Infopedia, etc) and other online services, such as online translators, in order to fully understand the meaning of tags. In some cases further research and analysis of other tags of a given user, or even a direct contact with this user by email was necessary in order to understand the exact meaning of a given tag.

The first step of tag analysis comprises grouping tag variants: a) language; b) simple/compound; c) abbreviations and acronyms; d) singular/plural; e) capital letter/small letter. Then a Key-tag is assigned to each of these groups according to the rules presented in section 3. Following, there are two examples of tags and their assigned key-tags:

- Tags: `metadados, metadata, meta-data, metadata/, métadonnées, metadata.tags`; Key-tag: `METADATA`;
- Tags: `informationscience, information science, information.science, Ciències de la informació, is`; Key-tag: `INFORMATION SCIENCE`;

The above key-tags show a variation in :

[2] Post-coordination is the principle by which the relationship between concepts is established at the moment of outlining a search strategy (Angulo Marcial, 1996 apud Menezes; Cunha; Heemann, 2004).

- spelling: information science, informationscience, information.science and is;
- form (Singular/Plural): metadata, metadados, métadonnées;
- language: information science (EN), ciènces de la informació (CA); metadados (PT), metadata (EN) and métadonnées (FR).

The examples above also show the two kinds of compound tags. Compound Tags focus/modifier like information science are assigned to only Key-tag. Tags composed of two focus components like metadata.tags are assigned to two distinct Key-tags: Metadata and Tags.

After Key-tags definition, an analysis to verify which DC Properties correspond to these tags is carried out. This analysis becomes more complex as the DCMI Terms definitions are purposely general enough so that the description of the electronic documents with a small, though sufficient, number of metadata is possible.

5. Complementary Properties - Results from the Pilot Study

In the pilot study it was analyzed data related to the first five resources of the data set. This implied the analysis of a total of 311 tags with 1141 occurrences and assigned by 355 users.

The accomplishment of the pilot study was also important in order to compare its results with the results of KoT. This study, is, however, much more detailed than the one in KoT which generated some indicative results: 1) "Users apply tags not only to describe the resource, but also to describe their relationship with them (e.g. to read, to print,…)"; 2) "Do tags correspond to atomic values? Many of the tags have more than one value, with potential results in more than one metadata element assigned"; 3) "Into which DC elements can tags be mapped? 14 out of the 16 DC elements, including Audience, have been allocated" (Baptista et al., 2007).

The results from KoT indicated that the following new elements could be added to the DC Social Tagging Application Profile: Action Towards Resource (e.g., to read, to print…); To Be Used In (e.g. work, class); Rate (e.g., very good, great idea) and Depth (e.g. overview).

The preliminary results from the current pilot study confirm the need for the proposal of new metadata elements for Social Tagging applications. However, it points out for some more elements than KoT did. The results of this study are presented in the following sections and, when pertinent, they will be compared with the results of KoT.

From the 311 tags analyzed in the pilot study, 212 Key-tags were created. From this amount, 159 Key-tags (75%) of which corresponded to the following DC properties: Creator, Date, Description, Format, Is Part Of, Publisher, Subject, Title and Type. From these, 90,5% corresponds to Subject. The other properties present the following percentages of allocation: Type 5%; Creator, Is Part Of and Title 3,1% each; Date and Publisher 1,3% each and Format 0,6%.

No DC properties could be assigned to the other 53 Key-tags (25%). New complementary properties were defined and their definition is still in process. The following properties that were identified in the pilot study will be described: Action, Category, Depth, Rate, User Name, Utility and Notes.

From these eight possible new properties, four had already been suggested in the KoT. Nonetheless, until the end of the full study, others may be added, or even, some of the ones proposed here may be withdrawn, depending on the evolution of the study.

In the group of the 53 Key-tags the following percentages for the properties proposed were observed: Action, Rate and Utility (15,1% each), Category (11,3%), Depth (9,4%), Notes (7,5%) and User Name (1,9%). There is also a group of Key-tags (24,5%) to which it was not possible to assign or propose any property as their meaning in relation to the resources and users was not possible to identify.

5.1. Action

There is a group of Key-tags that represents the action of the user in relation to the tagged resource. It is a kind of tag that can be easily identified since the action is expressed in the very term itself when tagging the resource. As example the tags which represent the action To Read, attributed to 6 users, all from Delicious: `_toread`, `a_lire`, `toread`.

5.2. Category

This property includes Tags whose function is to group the resources into categories, that is, to classify the resources. The classification is not determined by subjects or theme of the resource, since, in these cases, the key-tags could correspond to the Subject property.

This property is not easy to identify, since it is necessary to analyze the given tag in the context of the totality of tags that user has inserted, independently of the resource under analysis. In some cases it may become necessary to analyze the whole group of resources the user has tagged with the tag that is object of analysis.

For instance, during the analysis of the Key-tag `DC Tagged` it was noticed that the corresponding resources had also other tags with the prefix `dc:` (e.g.: `dc:contributor`, `dc:creator`, `dc:Publisher, dc:language` or `dc:identifier`, among others). It was concluded that the tag "`DC Tagged`" could be applied to group all the resources that were tagged by tags that were prefixed by `dc:`. Therefore it was considered a "Category" since it is not a classification of subjects or a description of the content of the resource.

5.3. Depth

This type of tag confers the degree of intellectual depth to the tagged resource. As Word Net, Depth "degree of psychological or intellectual profundity" (WorNet, 2008). A resource was tagged by six users who assigned the following tags to represent the degree of profundity of the resource: `diagram`, `doc/intro`, `overview`, `semanticweb.overview`, `semwebintro`. These tags mean that users are describing a resource which content is thought as a schematic or a summarized explanation, introductory and general.

5.4. Notes

This element may be proposed to represent the tags that are used as a note or reminder. As WordNet, "a brief written record" that has the objective of registering some observations concerning the resource, but that does not refer to its content and does not intend to be used as its classification or categorization (WordNet, 2008). A note should be understood as: an annotation to remind something; observation, comment or explanation inserted in a document to clarify a word or a certain part of the text (Infopedia, 2008).

From the five analyzed resources, the following tags considered as "Notes" were identified: `Hey`, `Ingenta`, `OR2007`, `PCB Journal Club`. For instance, there is a resource that received the tags `Hey` and `OR2007`. The first tag, `Hey`, refers to Tony Hey, a well-known researcher who made a debate on important issues that were related to the tagged resource[3].

The second tag makes reference to the Open Repositories 2007, event where Tony Hey mentioned above made a Keynote speech. However, interestingly enough, the tagged resource does not have any direct relation neither with that event nor with Tony Hey[4].

[3] This information was given by the user who assigned the tags.

[4] This information confirmed by the author of the resource himself (the creator).

5.5. Rate

Rate, meaning pattern, category, class or quality is important to include tags that are evaluating the tagged resource. Thus, the user categorizes the resource according to its quality when using this type of tag.

The following tags were related to the property: `academic`, `critical`, `important`, `old`, `great`, `good` and `vision`. These are generally easily identified as Rate in each one of the terms. In other cases, the tags may be doubtful and it becomes necessary to analyze them in relation to the tags assigned by the user to the resource under analysis as well as to the whole collection of resources tagged by that user. For instance, the tag `Vision` could have several meanings, but, after an analysis to the collection of resources, it may be concluded that it is classifying the quality of the resource.

5.6. User Name

The Tag "User Name" labels the resource with the name of a user. The analyzed resource had the name of the user of the tagged resource.

Only one tag of this type was identified in the pilot study. Despite the preliminary results presented here, it is assumed that here may be other occurrences.

5.7. Utility

This property would gather the tags that registered the utility of the resource for the user.

It represents a specific categorization of the tags, so that the user may recognize which resources are useful to him in relation to certain tasks and utilities.

`Maass` is a tag that was bundled in "Study". The term represents the name of a teacher, information found in the user's notes in two resources tagged with `Maass`: "Forschung von Prof. Maass an der Fakultat Digitale Medien an der HFU"; and "Unterlagen für Thema 'Folksonomies' für die Veranstaltung "Semantic Web" bei Prof. Maass".

6. Final Considerations

In the pilot study 212 key-tags were generated. DC properties could be assigned to 159 (75%) of those. The identified new properties were assigned to 40 key-tags (18,9%) and 13 key-tags (6,1%) were left without assignment because it was not possible to identify their meaning. As this data shows, DC properties can be assigned to a great part of the tags analyzed in the pilot study. However, still, 25% of them are left out.

The final study has already been finalized and although it is not yet possible to show the final results, it is possible to say that the percentage tags unassigned to DC elements is higher and it will probably range between 35% and 45% (39,5% is the provisory number, but some further analysis will still be done). It is not possible to assign properties to a great number of those tags because their meaning could not be identified. However, new properties could be assigned to most of them (the provisory number for tags assigned with new properties is 26,5%, while the provisory number for tags left unassigned is 13%).

DC plays a fundamental role as a foundation for metadata interoperability. From this study it is evident that DC keeps this role even in the presence of a paradigm shift, as withWeb 2.0 and the social tagging applications. However, as in these applications the user is in the centre of the description process, there is a significant number of new kinds of values (terms/tags) not previously foreseen in the scope of DC and to which current DC properties cannot be assigned.

This research aims at discovering if the DC properties have the necessary semantics to hold tags and, if not, it aims at finding which other properties that hold the lacking semantics can be coined to complement DC and to be used in social tagging applications. This application profile will allow rich descriptive tags to be handled by metadata interoperability protocols and consequently, to enrich the semantic Web.

This work begun with a pilot study for the first five resources of the KoT data set in order to refine the methodology and have a preliminary overview of the possible new properties that could be identified, if any. This article presents the results from the pilot study and already gives some lights on the final study. The final research results will then be submitted to the DC community for evaluation and validation purposes.

References

Baptista, Ana Alice, et al. (2007). Kinds of tags: progress report for the DC-social tagging community. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007*. Retrieved September 4, 2007, from http://hdl.handle.net/1822/6881.

Curras, Emília. (2005). *Ontologías, taxonomía y tesauros: manual de construcción y uso*. Gijón, Asturias: Ediciones Trea.

DCMI Usage Board. (2008). *DCMI metadata terms*. Retrieved March 10, 2008, from http://dublincore.org/documents/dcmi-terms/.

Guy, Marieke, and Emma Tonkin. (2006). Folksonomies: tidying up tags?. *D-Lib Magazine 12*(1). Retrieved December 12, 2006, from http://www.dlib.org/dlib/january06/guy/01guy.html.

INFOPEDIA. (2008). Retrieved March 10, 2008, from http://www.infopedia.pt.

Menezes, Esteria Muszkat, Miriam Vieira da Cunha, and Vivian Maria Heeman. (2004). *Glossário de análise documentaria*. São Paulo: ABECIN.

O'Reilly, Tim. (2005, October 1). Web 2.0: Compact definition?. Message posted to http://radar.oreilly.com/archives/2005/10/web_20_compact_definition.html.

Tonkin, Emma, et al. (2007). Kinds of tags: a collaborative research study on tag usage and structure (presentation). *The 6th European Networked Knowledge Organization Systems (NKOS) Workshop, at the 11th ECDL Conference, Budapest, Hungary.* Retrieved December 10, 2007, from http://www.cs.bris.ac.uk/Publications/pub_master.jsp?id=2000724.

Wal, Thomas Vander. (2006). *Folksonomy definition and wikipedia*. Retrieved November 22, 2006, from http://www.vanderwal.net/random/entrysel.php?blog=1750.

WORDNET. (2008). Retrieved November 22, 2006, from http://wordnet.princeton.edu/.

Full Papers

Session 2:

Semantic Integration, Linking, and KOS Methods

LCSH, SKOS and Linked Data

Ed Summers
Library of Congress, USA
edsu@loc.gov

Antoine Isaac
Vrije Universiteit Netherlands
aisaac@few.vu.nl

Clay Redding
Library of Congress, USA
cred@loc.gov

Dan Krech
Library of Congress, USA
eikeon@eikeon.com

Abstract

A technique for converting Library of Congress Subject Headings MARCXML to Simple Knowledge Organization System (SKOS) RDF is described. Strengths of the SKOS vocabulary are highlighted, as well as possible points for extension, and the integration of other semantic web vocabularies such as Dublin Core. An application for making the vocabulary available as linked-data on the Web is also described.

Keywords: metadata; semantic web; controlled vocabularies; SKOS; MARC; RDF; Dublin Core; identifiers

1. Introduction

Since 1902 the mission of the Cataloging Distribution Service at LC has been to enable libraries around the United States, and the world, to reuse and enhance bibliographic metadata. The cataloging of library materials typically involves two broad areas of activity: descriptive cataloging and subject cataloging. Descriptive cataloging involves the maintenance of a catalog of item descriptions. Subject cataloging on the other hand involves the maintenance of controlled vocabularies like the Library of Congress Subject Headings and classification systems (Library of Congress Classification) that are used in descriptive cataloging. As Harper (2007) has illustrated, there is great potential value in making vocabularies like LCSH generally available and reference-able on the Web using semantic web technologies.

The Library of Congress makes the Library of Congress Subject Headings (LCSH) available for computer processing as MARC, and more recently as MARCXML. The conventions described in the MARC21 Format for Authority Data are used to make 265,000 LCSH records available via the MARC Distribution Service. The Simple Knowledge Organization System (SKOS) is an RDF vocabulary for making thesauri, controlled vocabularies, subject headings and folksonomies available on the Web (Miles et al., 2008). This paper describes the conversion of LCSH/MARC to SKOS in detail, as well as an approach for making LCSH available with a web application. It concludes with some ideas for future enhancements and improvements to guide those who are interested in taking the approach further.

The remainder of this paper will use LCSH/MARC to refer to Library of Congress Subject Headings represented in machine-readable format using the MARCXML format; and LCSH/SKOS will refer to LCSH represented as SKOS. A basic understanding of RDF, SKOS and LCSH is assumed for understanding the content within.

2. Representing LCSH as SKOS

2.1. Basic Model

Harper (2006) has done significant earlier work imagining LCSH/MARC as SKOS, and has provided a concrete XSLT mapping for converting MARCXML authority data to SKOS. Both SKOS and LCSH/MARC have a concept-oriented model. LCSH/MARC gathers different forms of headings (authorized/non authorized) into records that correspond to more abstract conceptual entities, and to which semantic relationships and notes are attached. Similarly SKOS vocabularies are largely made up of instances of skos:Concept, which associate a “unit of thought” with a URI. SKOS concepts have lexical labels and documentation attached to them, and can also reference other concepts using a variety of semantic relationships.

2.2. Concepts

Since every MARC Authority record supplied by LC contains a Library of Congress Control Number (LCCN) in the 001 MARC field, it makes a good candidate for the identification of SKOS concepts. LCCNs are designed to be persistent, and are guaranteed to be unique. SKOS requires that URIs are used to identify instances of skos:Concept. Semantic Web technology—as specified by RDF (Frank Manola, et al., 2004) — and Linked Data practices also encourage the use of HTTP URLs to identify resources, so that resource representations can easily be obtained (Sauermann et al., 2007). Of course LCCNs are not URLs, so the LCCN is normalized and then incorporated into a URL using the following template http://lcsh.info/{lccn}#concept.

The use of the LCCN in concept URIs marks a slight departure from the approach described by Harper (2006), where the text of the authorized heading text was used to construct a URL: e.g. http://example.org/World+Wide+Web. The authors preferred using the LCCN in concept identifiers, because headings are in constant flux, while the LCCN for a record remains relatively constant. General web practice (Berners-Lee, 1998) and more specifically recent semantic web practice (Sauermann et al., 2007) encourage the use of URIs that are persistent, or change little over time. Persistence also allows metadata descriptions that incorporate LCSH/SKOS concepts to remain unchanged, since they reference the concept via a persistent URL.

2.3. Lexical Labels

The MARC21 Authority format distinguishes between authorized (1XX) and non-authorized (4XX) headings. Similarly the SKOS vocabulary provides two properties, skos:prefLabel and skos:altLabel, that that allow a concept to be associated with both preferred and alternate natural language labels. In general, this allows authorized and non-authorized LCSH headings to be mapped directly to skos:prefLabel and skos:altLabel properties in a straightforward fashion.

However, a significant amount of information is also lost. The specific MARC field used to represent an authorized heading captures the type of concept: chronological (148), topical (150), geographic (151), genre/form (155). It is important for the LCSH/SKOS representation to capture the notion of different types of concepts as well (see below).

Also, a number of LCSH/MARC authorized headings are the result of combining other headings, a technique that is commonly referred to as pre-coordination. For example, a topical heading Drama can be combined with the chronological heading 17th century, which results in an LCSH/MARC record with the authorized heading Drama--17th century. In LCSH/MARC this information is represented explicitly, with original headings and subdivision 'facets'. In the LCSH/SKOS representation, headings with subdivisions are flattened into a literal, e.g. “Drama--17th century”. This is an area where an extension of SKOS could be useful.

SKOS has been designed for use in a multi-lingual environment. SKOS users are encouraged to use language tags to identify the language of particular label (Isaac et al., 2008):

```
ex:animals rdf:type skos:Concept;
skos:prefLabel "animals"@en;
skos:prefLabel "animaux"@fr.
```

However, not all lexical labels in LCSH/SKOS are in English, e.g. Cueva de La Griega (Spain). Since this heading contains both Spanish and English it’s not entirely clear what language tag to use. In addition LCSH/MARC records do not contain an indicator of what languages are used in heading fields—so it would be challenging to programmatically assign them.

2.4. Semantic Relationships

LCSH/MARC uses the 5XX fields to link an authorized heading to other related authorized headings. SKOS provides a rich set of semantic relationships between conceptual resources, including: skos:related, skos:broader, skos:narrower.

The semantic relationships present in LCSH/MARC are easily translated into LCSH/SKOS. The links in LCSH/MARC use the established heading as references, whereas in LCSH/SKOS conceptual resources are linked together using their URIs. This requires that the conversion process lookup URIs for a given heading when creating links. In addition LCSH/MARC lacks narrower relationships, since they are inferred from the broader relationship. When creating skos:broader links, the conversion process also creates explicit skos:narrower properties as well. Once complete conceptual resources identified with URIs are explicitly linked together in a graph structure similar to Figure 1, which represents concepts related to the concept “World Wide Web”.

2.5. Documentation Properties

LCSH/MARC has a collection of fields that document aspects of the heading, including: general notes (667), source data (670), historical data (678), and examples (681). The SKOS vocabulary also includes documentation properties which can be used to represent LCSH/SKOS: skos:note, skos:editorialNote, skos:definition, skos:scopeNote, skos:changeNote, skos:historyNote. These properties are easily converted from LCSH/MARC to LCSH/SKOS, and require little massaging.

2.6. Using non-SKOS Documentation Properties

LCSH/MARC contains other features such as a relevant Library of Congress Classification Number ranges, the date that the record was created, and the date that a record was last modified. While the SKOS vocabulary itself lacks properties for capturing this information, the flexibility of RDF allows other vocabularies such as Dublin Core to be imported and mixed into SKOS descriptions: dcterms:lcc, dcterms:created, dcterms:modified. The flexibility to mix other vocabularies in to resource descriptions at will, without being restricted to a predefined schema is a powerfully attractive feature of RDF.

2.7. LCSH/SKOS Mapping

The general transformations above have been summarized into the following set of mappings.

MARC Field	Feature/Function	RDF Property	Value of the Property/Comments
010	Control Number	rdf:about	the URI for the skos:Concept instance
150	Topical Term	skos:prefLabel	subfields: a, b, v, x, y, z
151	Geographic Term	skos:prefLabel	subfields: a, b, v, x, y, z
450	See From Tracing (Topical Term)	skos:altLabel	subfields: a, b, v, x, y, z
451	See From Tracing (Geographic Name)	skos:altLabel	subfields: a, b, v, x, y, z
550	See Also From Tracing (Topical Term)	skos:broader	only use this property when subfield w is 'g'; use value to lookup Concept URI
550	See Also From Tracing (Topical Term)	skos:related	only use this property when subfield w is not present with 'g' or 'h' in position 0 ; use value to lookup Concept URI
551	See Also From Tracing (Geographic Name)	skos:broader	only use this property when subfield w is 'g'; use value to lookup Concept URI
551	See Also From Tracing (Geographic Name)	skos:related	only use this property when subfield w is not present with 'g' or 'h' in position 0 ; use value

MARC Field	Feature/Function	RDF Property	Value of the Property/Comments
			to lookup Concept URI
667	non public general note	skos:note	subfield: a
670	Source data found	dcterms:source	subfields: a, b, u
675	Source data not found	skos:editorialNote	subfield: a
678	Biographic or historical data	skos:definition	subfields: a, b, u
680	Public general note	skos:scopeNote	subfields: a,i
681	Subject example tracing note	skos:example	subfields: a, i
682	Deleted heading information	skos:changeNote	subfields: a, i
688	Application history note	skos:historyNote	subfield: a
008	Fixed Length Data Elements	dcterms:created	positions: 0-5
005	Date and time of last transaction	dcterms:modified	
053	LC Classification Number	dcterms:lcc	subfield: a

2.8. LCSH/SKOS Illustrated

Once a given LCSH/MARC record has been converted to LCSH/SKOS an RDF graph similar to Figure 1 has been created. Note: documentation properties have been left out for display purposes.

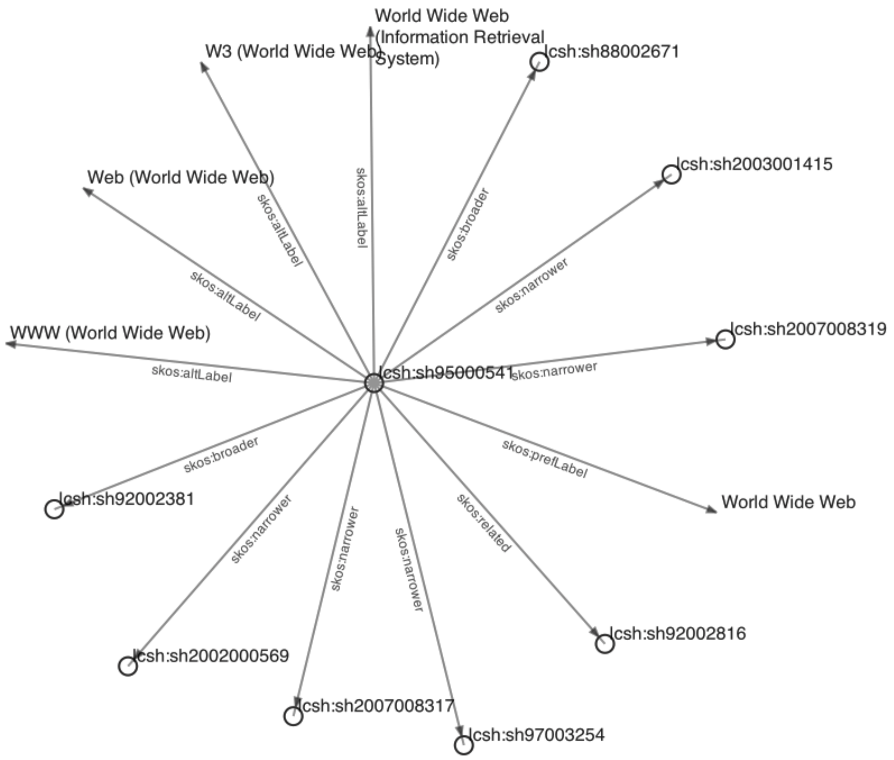


FIG. 1. SKOS Concept Graph

This example is for the concept “World Wide Web”. The textual links between LCSH/MARC records are made into explicit URI links between conceptual resources, as illustrated in Figure 2

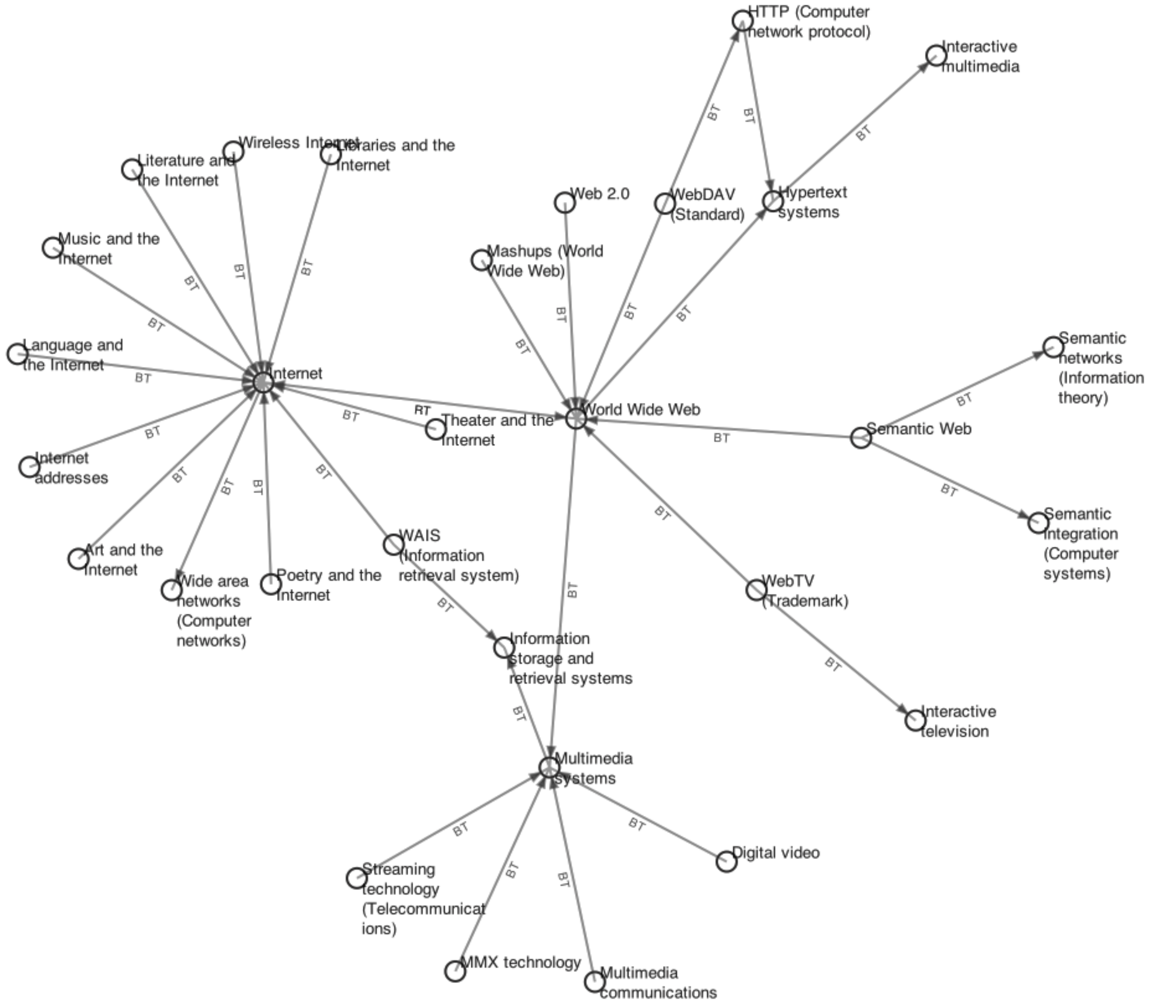

HTTP (Computer network protocol)
Interactive multimedia
Literature and the Internet
Wireless Internet
Libraries and the Internet
Web 2.0
WebDAV (Standard)
Hypertext systems
Music and the Internet
Mashups (World Wide Web)
Language and the Internet
Internet
Semantic networks (Information theory)
World Wide Web
Theater and the Internet
Semantic Web
Internet addresses
Semantic integration (Computer systems)
WAIS (Information retrieval system)
Art and the Internet
Wide area networks (Computer networks)
Poetry and the Internet
WebTV (Trademark)
Information storage and retrieval systems
Interactive television
Multimedia systems
Digital video
Streaming technology (Telecommunications)
MMX technology
Multimedia communications
BT

FIG. 2. Semantic Relationships between Concepts.

3. Delivering LCSH/SKOS as Linked Data

3.1. Cool URIs for LCSH/SKOS Concepts

Implicit in the translation of LCSH/MARC to LCSH/SKOS is the minting of hundreds of thousands of URIs for conceptual resources. It is a key aspect of the semantic web and linked data (Sauermann et al., 2008) that resources are identified with resolvable HTTP URLs. The notion of “following your nose” on the World Wide Web is what allows a distributed set of machine readable descriptions to be built. The Architecture of the World Wide Web (Jacobs et al., 2004) makes a distinction between URIs for *Information Resources* (descriptions of things) and URIs for *Non-Information Resources* (the things themselves). SKOS concepts (e.g. *Mathematics*) are clearly not available on the web, so special care must be taken in minting URIs for them. Sauermann (2008) provides specific guidance on how to make resources available on the semantic web. As described in 2.2, URLs of the pattern *http://lcsh.info/{lccn}#concept* are created for each LCSH/SKOS concept. The use of hash URIs for SKOS concept simplifies the web server implementation; since the server isn’t required to redirect using a *303 See Other* HTTP status code, when the URI for the concept is requested.

3.2. Content Negotiation

The authors chose to deliver multiple representations of LCSH/SKOS concepts on the Web using a technique called content-negotiation. When deciding what content to deliver to an HTTP client, a web server can examine the *Accept* header sent by the client, to determine the preferable representation of the resource to send (Berrueta et al, 2008). The LCSH/SKOS delivery application currently returns the following representations: rdf/xml, text/n3, application/xhtml+xml, application/json representations, using the URI patterns illustrated in Figure 3.

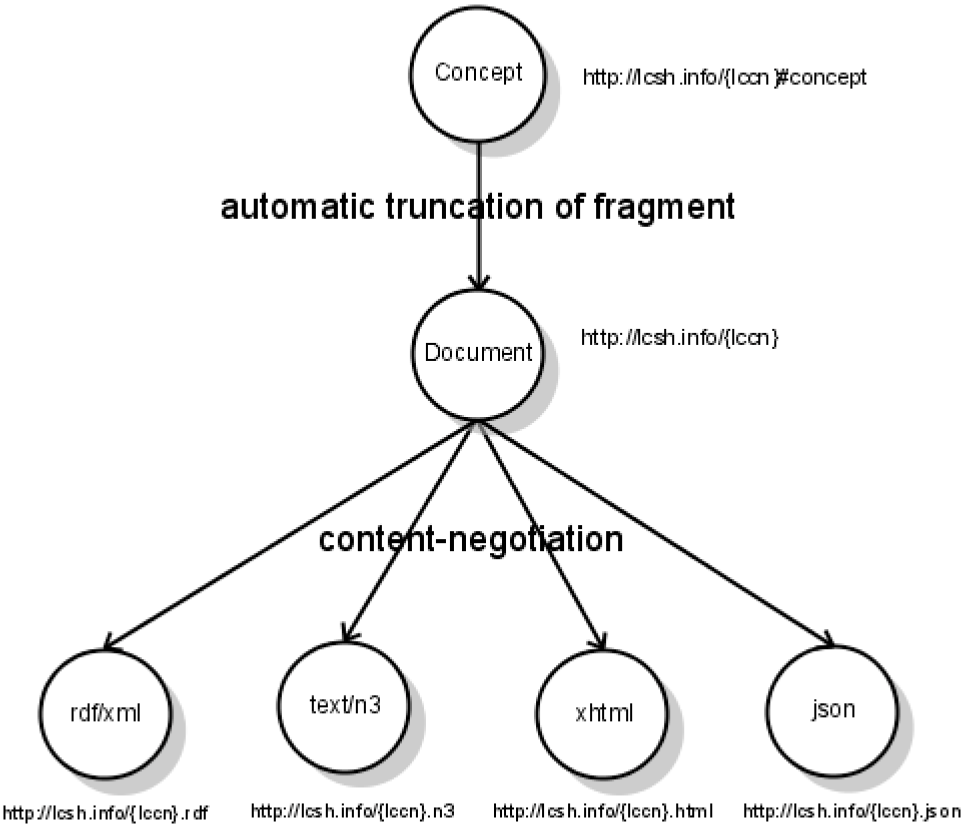
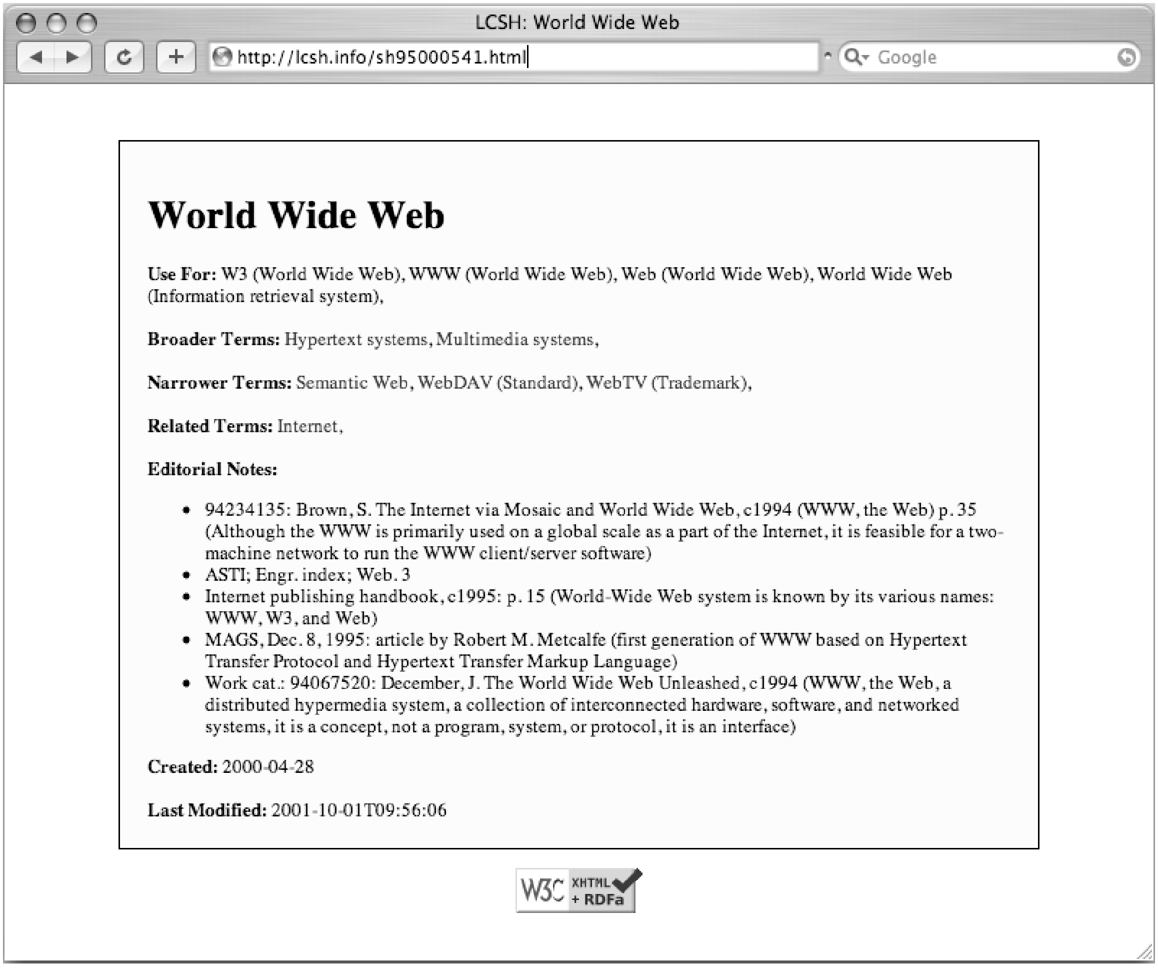

Concept
http://lcsh.info/{lccn}#concept
automatic truncation of fragment
Document
http://lcsh.info/{lccn}
content-negotiation
rdf/xml
text/n3
xhtml
json
http://lcsh.info/{lccn}.rdf
http://lcsh.info/{lccn}.n3
http://lcsh.info/{lccn}.html
http://lcsh.info/{lccn}.json

FIG. 3. URL Patterns.

The use of content-negotiation allows the LCSH/SKOS concept scheme to be browsed naturally by "following your nose" (Summers, 2008) to related concepts, simply by clicking on links in your browser (see Figure 4). It also allows semantic web and web2.0 clients to request machine-readable representations using the very same LCSH concept URIs. In addition the use of RDFa (Adida et al., 2008) allows browsers to auto-detect and extract semantic content from the human readable XHTML.



World Wide Web

Use For: W3 (World Wide Web), WWW (World Wide Web), Web (World Wide Web), World Wide Web (Information retrieval system),

Broader Terms: Hypertext systems, Multimedia systems,

Narrower Terms: Semantic Web, WebDAV (Standard), WebTV (Trademark),

Related Terms: Internet,

Editorial Notes:

- 94234135: Brown, S. The Internet via Mosaic and World Wide Web, c1994 (WWW, the Web) p. 35 (Although the WWW is primarily used on a global scale as a part of the Internet, it is feasible for a two-machine network to run the WWW client/server software)
- ASTI; Engr. index; Web. 3
- Internet publishing handbook, c1995: p. 15 (World-Wide Web system is known by its various names: WWW, W3, and Web)
- MAGS, Dec. 8, 1995: article by Robert M. Metcalfe (first generation of WWW based on Hypertext Transfer Protocol and Hypertext Transfer Markup Language)
- Work cat.: 94067520: December, J. The World Wide Web Unleashed, c1994 (WWW, the Web, a distributed hypermedia system, a collection of interconnected hardware, software, and networked systems, it is a concept, not a program, system, or protocol, it is an interface)

Created: 2000-04-28

Last Modified: 2001-10-01T09:56:06

FIG. 4. LCSH/SKOS Concept as RDFa XHTML.

4. Implementation Details

Remarkably little code (429 lines) needed to be written to perform the conversion and delivery of LCSH/SKOS. The Python programming language was used for both tasks, using a several open-source libraries:

- pymarc: for MARCXML processing (http://python.org/pypi/pymarc)
- rdflib: for RDF processing (http://python.org/pypi/rdflib)
- web.py: a lightweight web framework (http://python.org/pypi/web.py)
- webob: HTTP request/response objects, with content-negotiation support (http://python.org/pypi/WebOb)

The general approach taken in the conversion from LCSH/MARC to LCSH/SKOS differs somewhat from that taken by Harper (2006). Instead of using XSLT to transform records, the pymarc library was used, which provides an object-oriented, streaming interface to MARCXML records. In addition a relational database was not used, and instead the rdflib BerkeleyDB triple-store backend was used to store and query the 2,625,020 triples that make up the complete LCSH/SKOS dataset. The conversion process itself runs in two passes: the first to create the concepts and mint their URIs, and the second to link them together. To convert the entire dataset (377 MB) it takes roughly 2 hours, on a Intel Pentium 4 CPU 3.00GHz machine.

Readers interested in running the conversion utilities and/or the web application can check out the code using the Bazaar revision control system (http://bazaar-vcs.org) from http://inkdroid.org/bzr/lcsh.

5. Improvements and Future Directions

5.1. Extending SKOS

Since SKOS was designed as a general tool for knowledge organization systems (thesauri, classification schemes, subject heading lists, taxonomies, folksonomies) it lacks specialized features to represent some of the details found in LCSH/MARC. As discussed above in 2.3, LCSH/MARC distinguishes between several types of concepts: geographic, topical, genre/form, and chronological. However LCSH/SKOS has only one type of entity *skos:Concept* to represent all of these. As an RDF vocabulary, SKOS could easily be extended with new sub-classes of *skos:Concept: lcsh:TopicalConcept, lcsh:GeographicConcept, lcsh:GenreConcept,* and *lcsh:ChronologicalConcept*.

In addition LCSH/MARC uses pre-coordination to assemble authorized subject headings from the combination of other headings. These pre-coordinations use a variety of subfields to capture the type of facet used in a heading. Unfortunately this information is lost in SKOS since the *skos:prefLabel* property has for its range, and joins the subfields together with a '—'. Users of LCSH/SKOS will undoubtedly want to be able to identify the components of pre-coordinated concepts. Some LCSH/MARC records represent authorized subfield headings (180, 181, 182, 185), which were ignored by our initial conversion routine. It would be useful to represent these concepts using a SKOS extension. SKOS currently has an open issue (Miles, 2007) to explore how to represent coordinated concepts in SKOS, or to provide an extension pattern. Once a clear path is presented it would be useful to implement the solution in LCSH/SKOS.

5.2. Linking Open Data

One of the advantages of the semantic web and linked data is that traditionally isolated data sets can be integrated. Bizer (2007) provides guidance on how to link together semantic web resources using a variety of techniques. The LCSH/SKOS dataset has multiple places where links could be created to external datasets, including:

- GeoNames (http:///geonames.org) and the CIA World Fact Book (http://www4.wiwiss.fu-berlin.de/factbook/) for geographic headings.
- the RDF BookMashup (http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/) for links to items that prompted a LCSH concept to be created.
- dbpedia (http://dbpedia.org)

Furthermore, there are additional vocabularies at the Library of Congress such as the Library of Congress Classification, Name Authority File, and LCCN Permalink Service which could be made available as RDF. The authors are also involved in the conversion of the RAMEAU, a controlled vocabulary that is very similar to LCSH. Once converted these vocabularies would be useful for interlinking with LCSH.

5.3. Server Log Analysis

Even before being announced the LCSH/SKOS web application received thousands of hits a day from web-crawling robots (Yahoo, Microsoft. Google) and semantic web applications like Zitgist and OpenLink. The server logging was adapted to also capture accept HTTP header information, in addition to referrer, user agent, IP address, concept URI. After 6 months has elapsed it will be useful to review how robots and humans are using the site: the representations that are being received, how concepts are turning up search engines like Google, Yahoo, Swoogle (http://swoogle.umbc.edu/) and Sindice (http://sindice.com).

5.4. Discovery with SPARQL

The LCSH/SKOS web application makes the entire dataset of 2,625,020 RDF assertions available in a single file. This dump is useful for developers who want to be able to link up their data with LCSH/SKOS concept URIs. However, given the volume of data, a SPARQL endpoint

(Prud'hommeaux et al., 2008) would enable users to programmatically discover concepts without having to download and index the entire data set themselves. For example MARC bibliographic data has no notion of the LCCN for subjects that are used in descriptions. This indirection makes it impossible to determine which SKOS/LCSH concept URI to use without looking for the concept that has a given skos:prefLabel. A SPARQL service would make this sort of lookup trivial.

6. Conclusion

The conversion and delivery of Library of Congress Subject Headings as SKOS has been valuable on a variety of levels. The experiment highlighted the areas where SKOS and semantic web technologies excel: the identification and interlinking of resources; the reuse and mix-ability of vocabularies like SKOS and Dublin Core; the ability to extend existing vocabularies where generalized vocabularies are lacking. Hopefully the Library of Congress’ mission to provide data services to the library community will provide fertile ground for testing out some of the key ideas of semantic web technologies that have been growing and maturing in the past decade.

7. References

Adida, Ben, Mark Birbeck. (2008). *RDFa primer: Bridging the human and data webs.* Retrieved June 20, 2008 from http://www.w3.org/TR/xhtml-rdfa-primer/.

Berners-Lee, Tim. (1998). *Cool URIs don’t change.* Retrieved June 20, 2008 from http://www.w3.org/Provider/Style/URI.

Berrueta, Diego, Jon Phipps. (2008). *Best practice recipes for publishing RDF vocabularies*. W3C Working Draft. Retrieved June 20, 2008 from http://www.w3.org/TR/swbp-vocab-pub/.

Bizer, Chris, Richard Cyganiak and Tom Heath. (2007) *How to publish linked data on the web.* Retrieved June 20, 2008 from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

Harper, Corey. (2006). Authority control for the semantic web. Encoding Library of Congress subject headings. *International Conference on Dublin Core and Metadata Applications, Manzanillo, Mexico.* Retrieved June 20, 2008 from http://hdl.handle.net/1794/3268.

Harper, Corey, and Barbara Tillett. (2007). Library of Congress controlled vocabularies and their application to the semantic web*. Cataloging and Classification Quarterly, 43*(3/4). Retrieved June 20, 2008 from https://scholarsbank.uoregon.edu/dspace/handle/1794/3269.

Isaac, Antoine, and Ed Summers. (2008). *SKOS Simple Knowledge Organization System Primer.* Retrieved June 20, 2008 from http://www.w3.org/TR/skos-primer/.

Jacobs, Ian, and Norman Walsh. (2004). *Architecture of the world wide web, volume 1*. Retrieved June 20, 2008 from http://www.w3.org/TR/webarch/.

Manola, Frank, and Eric Miller. (2004). *RDF primer.* Retrieved June 20, 2008 from http://www.w3.org/TR/rdf-primer/.

Miles, Alistair. (2007). *Issue-40 concept coordination.* Retrieved June 20, 2008 from http://www.w3.org/2006/07/SWD/track/issues/40.

Miles, Alistair, and Sean Bechhofer. (2008). *SKOS Simple Knowledge Organization System Reference*. Retrieved June 20, 2008 from http://www.w3.org/TR/skos-reference/.

Prud'hommeaux, Eric, and Andy Seaborne. (2008). *SPARQL query language for RDF.* Retrieved June 20, 2008 from http://www.w3.org/TR/rdf-sparql-query/.

Sauermann, Leo, and Richard Cyganiak. (2007). *Cool URIs for the semantic web*. Retrieved June 20, 2008 from http://www.w3.org/TR/cooluris/.

Summers, Ed. (2008). Following your nose to the Web of Data*. Information Standards Quarterly, 20*(1). Retrieved June 20, 2008 from http://inkdroid.org/journal/following-your-nose-to-the-web-of-data.

Theme Creation for Digital Collections

Xia Lin
Drexel University
Philadelphia, PA, USA
xlin@drexel.edu

Jiexun Li
Drexel University
Philadelphia, PA, USA
jiexun.li@drexel.edu

Xiaohua Zhou
Drexel University
Philadelphia, PA, USA
xiaohua.zhou@drexel.edu

Abstract

This paper presents an approach for integrating multiple sources of semantics for the creating metadata. A new framework is proposed to define topics and themes with both manually and automatically generated terms. The automatically generated terms include: terms from a semantic analysis of the collections and terms from previous user’s queries. An interface is developed to facilitate the creation and use of such topics and themes for metadata creation. The framework and the interface promote human-computer collaboration in metadata creation. Several principles underlying such approach are also discussed.

Keywords: metadata creations; metadata authoring tools; topics and themes; human-computer collaboration

1. Introduction

The Internet Public Library (IPL: http://www.ipl.org) is one of the oldest digital libraries that is still actively maintained and used. Supported by a consortium of LIS schools (The IPL Consortium, n.d.), the IPL holds multiple collections of thousands of authoritative websites on various subjects. Most of these collections include sets of metadata records created by volunteering LIS students. Searching and browsing IPL collections are based on the metadata database. Because of this, it has been a priority for the IPL to create and maintain high-quality metadata within its current setting. The metadata will continue to be created by LIS students to support its mission as a teaching and learning environment for the Consortium member schools, yet high-quality metadata must be maintained to support its service to the public. It is essential for the IPL to have a powerful metadata creation tool that can be easily learned and used by professionals (or quasi-professionals) to create high-quality metadata for the digital library.

The objective of this research is to investigate how to incorporate multiple semantic sources to enhance metadata creation. Current IPL metadata consist of a set of well documented fields such as title, abstract, keywords, and subject headings. The subject headings are not a formally defined thesaurus but a set of loosely developed category terms. While the subject headings present a simple hierarchical view to the IPL collections, they do not provide strong associative and semantic relations among the headings and collections that a good thesaurus would otherwise provide. Thus, we sought solutions to build additional semantic relations among keywords, subject headings, topics and digital resources (Web pages) to enhance the IPL metadata. We particularly explored how context might be employed for metadata and how the context information might be extracted from both the semantic analysis of digital collections and the analysis of user’s search logs.

The remainder of this paper is organized as follows. First, we discuss various semantic sources for metadata. We then define topics and themes and introduce a framework for metadata subject representation using multiple semantics sources. In particular, we describe the language model for semantic mapping and a bottom-up procedure of theme creation. Finally, we introduce a system we developed to integrate all the semantics sources into one rich interface.

2. Metadata and Semantics

In computational linguistics, semantics refers to the relation between the words and sentences of a language and their meanings (Saeed, 2003). It is hypothesized that semantics can be extracted through lexical or statistical analysis of language and its structures. The meanings then can be represented by the data and structures obtained through the analysis. Similarly, semantics of metadata can be considered as the relation between metadata records and the content they represent. Metadata records are essentially "data + structures" that describe and represent various features of digital objects, including their content, context, and structures (Gill, et al. 1998). The semantics of metadata come from multiple sources. The first is the metadata standard. A metadata standard represents a consensus of how a specific type of digital objects should be described structurally. It provides a schema that specifies the metadata's namespaces, formats, required elements and allowable attributes, etc. Through naming and structuring the metadata elements, each standard provides a semantic framework that the user can "fill-in" values to create metadata records. The second source of the semantics comes from the metadata creation process. Typically, the standards do not give details on how a metadata record should be created. It is up to the metadata creator (most likely a human being) who interprets the content of the resource to be described and selects terms most appropriate for each entry of the metadata record. The human intelligence in this process provides the most significant semantic associations to connect metadata records to the content. The third semantic source of metadata is the semantics of the language. In particular, when a controlled vocabulary is used to create metadata records, the rich semantic relationships established within the controlled vocabulary enrich the semantics of metadata significantly.

There are many other semantic sources that have not been considered and incorporated into current metadata practice. One that seems to be obvious is the computerized semantic analysis of terms in a text collection. The semantic analysis can extract rich semantic relationships of terms over the whole collection to form "semantic metadata" (Haase, 2004). While such semantic metadata is still a lack of precision, incorporating selected terms from the list to enhance the standard-based metadata was considered a practical trend (Al-Khalifa, 2006).

Another useful source of semantics is the user's terms and usage patterns collected over time. How users search and interact with digital collections can provide valuable semantics for metadata creation. It could be an iterative process to improve the metadata with usage statistics. The more users use the collections, the more usage patterns will be collected and the better the metadata would be when the patterns are used appropriately.

Different sources can capture the semantics of a collection from a different perspective. It would be useful to integrate multiple semantic sources abovementioned to enhance the metadata creation process. In this research, we attempted to develop a framework and an authoring tool that would incorporate semantic mapping and usage patterns as semantics sources for metadata creation.

3. A Hierarchical Framework for Subject Representation in Metadata

3.1. Topics, Themes, and the Framework

Topic Maps (ISO13250, 2002) provide a new approach to represent knowledge and create associations among subjects and digital resources. As an established standard technology that includes well defined syntaxes, structures and the underlying reference model, Topic Maps describe knowledge structures through topics, associations, and occurrences in a formal model (Pepper, 2000). In this model, a theme is also defined as "a member of the set of topics comprising a scope within which a topic characteristic assignment is valid." The theme here is only used to define scopes for topics. However, as Pepper & Gronmo (2002) pointed out, both scopes and themes are the means to "putting context into topic maps."

We believe that there is a potential to expand the role of themes. A theme should be considered as a special topic or as "a topic in context." It can be used to provide contextual links to topics. It can become a higher level of subject indicators along with keywords, subject headings, and topics. In this research, we simply view a Topic as a subject with a name and multiple slots of properties. The properties may include different types of keywords generated from multiple sources either manually or automatically. Then, we view a Theme as a special type of topics that unites several topics around a theme. Unlike in Topic Maps where the center of the universe is "topic," we are exploring to have the "theme" as the main unit that can have its own descriptive metadata and let topics to be "characteristics" of the theme. Our assumption is that users would be more interested in "topics in context" than topics. When a searcher sends a query to a collection, the searcher will likely be more satisfied to retrieve themes that provide specific topics and resources relevant to the query than to retrieve only the resources that match the user's query directly.

Figure 1 illustrates the hierarchical framework for subject representation in metadata. As mentioned above, the properties to describe topics and themes can be from multiple sources. In this study, we only include three sets: a set of topic signatures automatically generated from a collection, a set of keywords manually assigned, and a set of keywords identified and selected from the previous user's query terms. Details on topic signature generation will be introduced in Section 4. Each set of keywords can have different weights for the purpose of indexing and retrieval. The topic can also include properties of relevant resources (occurrences), either manually selected or automatically retrieved and inserted by search engines. For the IPL, two types of resource URLs are included as properties of a topic: URLs within the IPL and those outside the IPL.

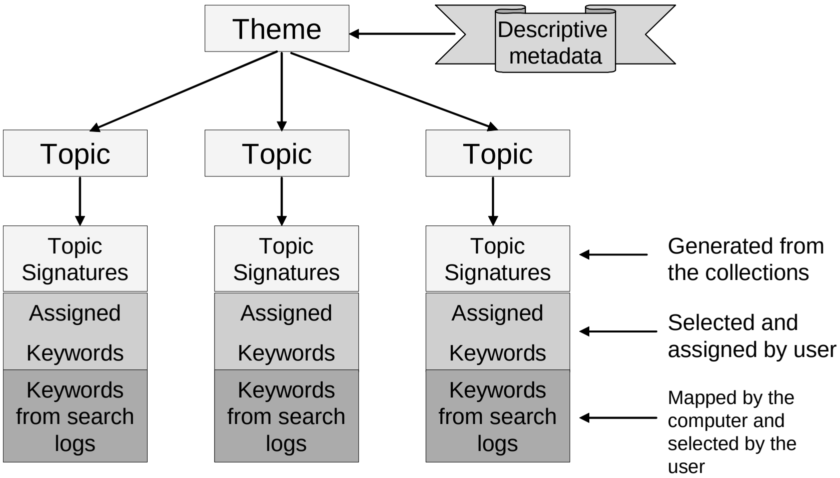

Theme
Descriptive metadata
Topic
Topic
Topic
Topic Signatures
Assigned Keywords
Keywords from search logs
Topic Signatures
Assigned Keywords
Keywords from search logs
Topic Signatures
Assigned Keywords
Keywords from search logs
Generated from the collections
Selected and assigned by user
Mapped by the computer and selected by the user

FIG. 1. A hierarchical framework for subject representation in metadata.

3.2. Collection-based Topic Signatures and Semantic Profiles

In the proposed framework of subject representation, topic signatures are a key concept. The topic signatures can be automatically generated through a topic signature model we developed (Zhou et al. 2006). The model is based on semantic mapping through a language modeling approach and a context-sensitive semantic smoothing method (Zhou et al., 2007a). Two types of mappings are created in this language model (Figure 2). One is called topic signatures that map from any term, w, in the collection to a list of topics, t's (represented by keywords, subject headings, or other indexing terms). The other is called semantic profiles that map a specific topic (t) to a set of terms (w's) that are most likely to co-occur with t in the collection, C.

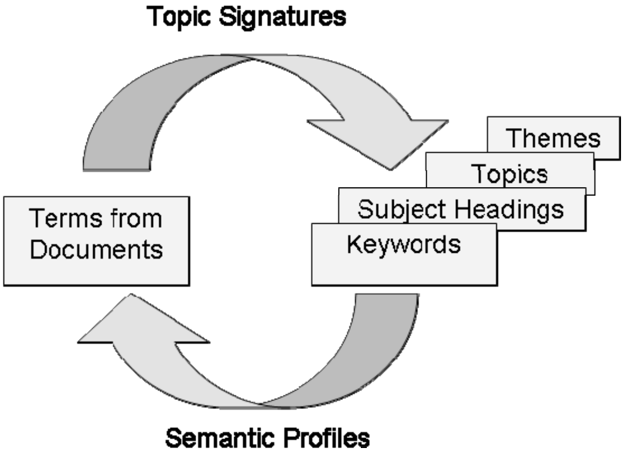


FIG. 2. Two types of semantic mapping: topic signatures and semantic profiles.

To create semantic profiles for keywords in a collection (*C)*, we first index all documents with individual terms and topics. For each topic t_k, we approximate its semantic profile using the terms *w*'s in the document set D_k containing t_k, ranked in the descending order of the conditional probability $p(w \mid t_k, C)$. We assume that the terms appearing in D_k are generated by a mixture model:

$$p(w \mid t_k, C) = (1-\alpha)p(w \mid t_k) + \alpha p(w \mid C)$$

where $p(w \mid t_k)$ is a topic model that represents the conditional probability of term *w* co-occurring with topic t_k. $p(w \mid C)$ is a background model describing the global distribution of terms in the collection *C*, and α accounts for the background noise. Not only does this mixture model capture the semantic associations between topics and terms in the topic model, but it also takes into account the overall term distribution of a collection in the background model. The model for t_k can be estimated using an expectation-maximization (EM) algorithm. Details of the model can be seen in (Zhou et al., 2006; Zhou et al., 2007a).

It is noted that the model represents an effective semantic mapping based not only on the content but also on the context. Due to the different focuses of different collections, the metadata used to describe the same terms or objects may vary from collection to collection. Our language model is able to capture the different semantic associations among topic signatures in different collections. Table 1 shows two topic signatures for the same topics in two different collections. For example, we conducted the semantic mapping on two of IPL collections: the IPL general collection and the IPL collection for Youth and Teens. The mapping results show strong "context interpretations." For example, the topic "reading" in the general (adult) collection is closely associated with "classics," "review," "humor," "literary," etc., while the same topic in the collection for teens is closely associated with "children's literature," "stories," "folklore," etc. Similarly, the topic signatures show that for the health issue, adults are more concerned about "mental health," "health care," "disease," etc. and the teens are more concerned about "fitness," "exercise," "nutrition," and "stress." When creating metadata for different collections, such suggestions of collection-based related terms would be very useful.

TABLE. 1. Examples of automatically-extracted topic signatures in different collections.

Topic	**Reading**		**Health**	
Collection	*IPL*	*IPL-Teens*	*IPL*	*IPL-Teens*
Topic signatures	**reading** classics review humor literary comic Cartoons Comics Censorship Rules FOIA biology Manga Native Americans Insurance Scientists Indian ……	**reading** children's literature stories folklore magazine story books author biographies social studies children instruments fantasy Games paleontology biography	**health** mental health health care disease activism disorders public Health psychology mental Illness reproduction medicine Safety Therapy prevention Pregnancy medical Nutrition Drugs ……	**health** fitness exercise nutrition stress tic panic medicine attention deficit ADHD depression add Teaching teens

Furthermore, for each keyword, the language model creates a semantic profile by mapping the keyword to a set of related terms in the specific collection. For example, as shown in TABLE 2, in the IPL Teens collection, the semantic profile of the keyword "stress" contains a list of highly associated words, including "depression," "anxiety," "health," "disorder," "eat," etc. The number attached to each term gives $p(w \mid \theta_{t_k}, C)$, the conditional probability of term *w* co-occurring with keyword t_k in collection *C*. Such a semantic profile can help us better understand the keyword in a particular context, and further decide whether to include the keyword in the metadata.

TABLE 2. An example of a semantic profile for the term "stress" in IPL Teens Collection.

Semantic profile	**Probability**
stress	0.0591
depress	0.0339
teen	0.0292
anxiety	0.0286
health	0.0284
disorder	0.0264
eat	0.0226
mental	0.0221
…	…

3.3. A Theme Creation Procedure

To create themes for the IPL, we developed a bottom-up procedure of theme creation and tested through a group of volunteers and students in selected classes. The process starts with examining users' search logs, reviewing the suggested topic signatures, and further identifying topics and themes as metadata. Figure 3 shows an example of instructions given to the theme creator. Figure 4 shows some examples of themes created by students following the procedures. Notice the themes are specific to the IPL collections and IPL users. Collectively, they indicate the content of the IPL from users' perspectives.

A. **Explore users needs.** You have access to the list of top 1000 query terms that the users used to search IPL most frequently in the past three months.

1) Browse through the list first.
2) Select terms to form groups as potential topics.
3) Identify some recurrent themes in the groups you identified.
4) Decide a theme you would like to work on

B. **Explore the collections**.

1) Use the topic signature and semantic profile tool to explore and collect relevant terms for the topics.
2) Use IPL search engines to identify relevant resources to the topics.
3) Use Google or other tools to identify relevant resources outside IPL.
4) Check if there are any pathfinders or spotlight within IPL that are relevant to the topics (both are related IPL finding-aids).

C. Create the theme

1) Describe the theme using the given metadata schema (i.e., complete the title, description, subject headings, and other descriptive fields).
2) Identify topics associated with the theme and generate the topic signatures.
3) Go over the automatically generated topic signatures and select the most relevant terms as keywords. Add your own terms as necessary.
4) You may need to go back to step A and B to complete the process.

FIG. 3. A procedure of theme creation for IPL collections.

- **Literature**
  - American Authors
  - American Literature
  - Banned Books
  - Shakespeare
- **History**
  - History of Military Conflicts
  - The American Revolution
  - Presidents of the United States of America
- **People**
  - Influential Americans
  - American Leaders
  - U.S. Presidents in Context
- **Teens**
  - Raising and Nurturing a Teenager
  - Teen Entertainment
- **Social Issues**
  - Race & Ethnicity
  - Public Policy Issues
  - Internet Popular Culture
- **Sciences**
  - Science Fair Projects
  - Topics in Science
  - Grade school research project
- **Technology**
  - Computers and Libraries
  - Personal Internet Entertainment
  - Web Service Hubs
  - Emerging Technologies
  - Entertainment/Social Networking
  - Cars
- **Environment**
  - The World We Live In
  - Environmentalism
  - Geography and World Locations
- **Health**
  - Physical Fitness
  - Your Body, Your Mind, Your Health
  - Diabetes prevention
  - Health Disorders and Prevention

FIG. 4. Sample titles of themes created for IPL collections.

The procedure and the process of theme creation highlight several principles we are developing and testing:

- Themes and topics can be created through an integrated process of both manual and automatic processes. It seems that the manual process could focus on the higher levels and the automatic process on the detailed and lower levels of the subject representations.

Each higher level of representations provides or enhances associative relationships of the lower level ones.

- Themes and topics can be represented dynamically based on the semantic analysis of the collections to be represented and on the analysis of previous user's interactions with the collections (for example, the search logs analysis). As the collections and the user's needs change over time, the meaning and the representation of the themes and topics may change dynamically.
- Context-sensitive themes do not need to be defined uniquely. Different users or systems can define same themes from different perspectives and with different names or sets of properties. Their similarities can be measured by their sharing properties (keywords and URLs, etc.). Similar themes will likely show a high degree of similarities.
- Themes and topics can be used to index and describe digital resources in multiple levels. The retrieval process can take advantages of such multi-level representations to provide search results in different granularity.

4. Implementation

To apply the framework and the procedure, we are developing an integrated system called "IPL KnowledgeStore." This section introduces the tools we adopted and the major features available in the rich interface we developed.

4.1. Semantic Mapping using Dragon Toolkit

While several complex natural language processing and statistical algorithms are needed to generate topic signatures and semantic profiles, Zhou et al. (2007b) also developed an open source toolkit to facilitate the mapping process. The toolkit, called the Dragon Toolkit, is a Java-based development package for language modeling and information retrieval, including text classification, text clustering, text summarization, and topic modeling. The toolkit is freely available for academic use at http://www.dragontoolkit.org. It provides many tools to map text collections with various representation schemes including words, phrases, ontology-based concepts and relationships. Specifically, in this research, we make use of the Dragon Toolkit APIs for the semantic mapping between terms and keywords (i.e., topic signatures and semantic profiles) in different collections.

4.2. An Integrated Interface

We are developing the IPL KnowledgeStore system using Adobe's Rich Internet Application (RIA) development environment, FLEX 3. Figure 5 shows a sample interface of the application. The interface functions as a "semantic aggregator" and a collaborative authoring workspace that provides access to multiple semantics sources, including the IPL metadata, topic signatures, semantic profiles created by the Dragon Toolkit, and the list of most frequently used search terms. The tool allows the user to create multiple types of digital objects such as subject terms, topics, themes, and metadata for IPL resources; each object itself is also described by a metadata. The user can create, modify, retrieve, and save these objects to a database or to XML files based on defined schemes. The interface provides rich interactive functions and links. Each source of semantics can be used separately or linked together. When a term in the center work space is selected, both sides of mappings (from the resource collections and from the user's terms) will be done automatically. Such mappings allow better use of associations hidden in the collections and in the user's interactions with the collections. The user can easily drag-and-drag terms from one slot to another and edit or select automatically generated terms to enhance the representations.

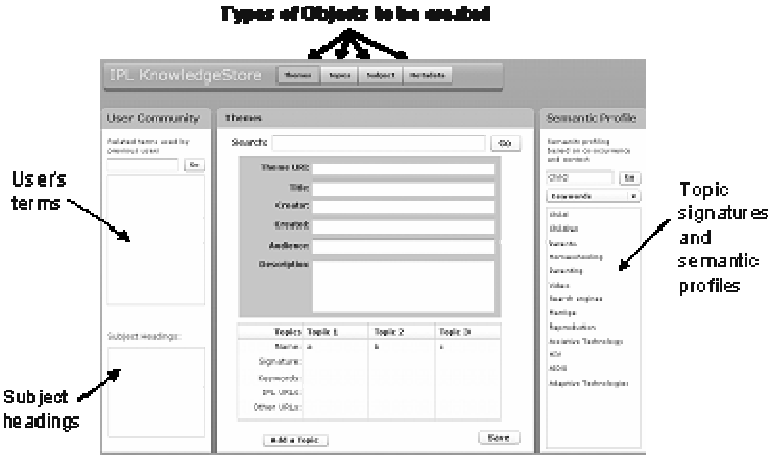


FIG. 5. A sample screen of the integrated interface.

5. Conclusions

Semantics of metadata might be considered as the relation between metadata records and the content they represent. In this paper, we examined two important considerations to enhance the semantics. One consideration is how to enhance content representation with context, and the other is how to integrate multiple sources of semantic sources for the purposes of metadata creation. We showed that, topics and themes could be created with a combination of automatic semantic mapping and human interpretation. The semantic mapping utilizes both content and context when suggesting topic signatures and semantic profiles for subject terms. The human users can take the suggestions to create topical themes or metadata with additional association and context interpretations. We also developed an integrated interface for metadata creation. The interface allows users to select subject terms from various semantic sources, including terms used in existing metadata records, terms from the subject headings, terms from topic signatures and semantic profiles created by the automatic semantic mapping, and terms from the search logs.

Much more research needs to be done to address the need for semantic enrichment of metadata. We plan to continue developing our system to examine several principles discussed in this paper and test the effectiveness and usability of the interface as a metadata authoring tool for the IPL. Finally, we hope to further refine the concepts of topics and themes and use them for multi-level subject indexing (such as keyword indexing, topic indexing, and theme indexing) for metadata and digital collections.

References

Al-Khalifa, Hend S., and Hugh C. Davis. (2006). The evolution of metadata from standards to semantics in e-Learning applications. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark, 2006,* (pp. 69-72).

Cruse, Alan. (2004). *Meaning in language: An introduction to semantics and pragmatics* (2nd ed.). Oxford: Oxford University Press.

DCMI. (2008). *Dublin Core Metadata Element Set, version 1.1.* Retrieved March 30, 2008, from http://dublincore.org/documents/dces/.

Gill, Tony, Anne Gilliland, and Murtha Baca. (1998). *Introduction to metadata: pathways to digital information*. Los Angeles, California: J. Paul Getty Trust. Retrieved March 30, 2008 from http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html.

Haase, Kenneth. (2004). Context for semantic metadata. *Proceedings of the 12th Annual ACM international Conference on Multimedia, New York, USA, 2004,* (pp. 204-211). Retrieved from http://doi.acm.org/10.1145/1027527.1027574.

The IPL Consortium. (n.d.) *The Internet Public Library*. Retrieved March 30, 2008, from http://www.ipl.org.

ISO/IEC (2002). *ISO/IEC 13250 -Topic Maps* (2nd ed.). Retrieved March 30, 2008, from, http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf.

Manning, Christopher D., and Hinrich Schütze. (1999). *Foundations of statistical natural language processing.* Cambridge, Mass: MIT Press.

Pepper, Steve. (2000). *The TAO of Topic Maps.* Retrieved March 30, 2008 from http://www.ontopia.net/topicmaps/materials/tao.html.

Pepper, Steve, and Grønmo, Geir Ove. (2002). *Towards a general theory of Scope*. Retrieved March 30, 2008, from http://www.ontopia.net/topicmaps/materials/scope.htm.

Saeed, John I. (2003). *Semantics*. *Introducing liguistics, 2*. Malden. MA: Blackwell Pub.

Zhou, Xiaohua, Xiaohua Hu, Xiaodan Zhang, Xia Lin, and Il-Yeol Song. (2006). Context-sensitive semantic smoothing for the language modeling approach to genomic IR. *Proceedings of the 29th Annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington, USA, 2006,* (pp. 170-177). Retrieved from http://doi.acm.org/10.1145/1148170.1148203.

Zhou, Xiaohua, Xiaohua Hu, and Xiaodan Zhang. (2007a). Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering, 19*(9), 1276-1287.

Zhou, Xiaohua, Xiaohua Zhang, and Xiaodan Hu. (2007b). Dragon Toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007, Patras, Greece* (pp. 197-201).

Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain

Boris Lauser
FAO, Italy
boris.lauser@fao.org

Gudrun Johannsen
FAO, Italy
gudrun.johannsen@fao.org

Caterina Caracciolo
FAO, Italy
caterina.caracciolo@fao.org

Willem Robert van Hage
TNO Science & Industry /
Vrije Universiteit
Amsterdam, the Netherlands
wrvhage@few.vu.nl

Johannes Keizer
FAO, Italy
johannes.keizer@fao.org

Philipp Mayr
GESIS Social Science
Information Centre
Bonn, Germany
philipp.mayr@gesis.org

Abstract

Knowledge organization systems (KOS), like thesauri and other controlled vocabularies, are used to provide subject access to information systems across the web. Due to the heterogeneity of these systems, mapping between vocabularies becomes crucial for retrieving relevant information. However, mapping thesauri is a laborious task, and thus big efforts are being made to automate the mapping process. This paper examines two mapping approaches involving the agricultural thesaurus AGROVOC, one machine-created and one human created. We are addressing the basic question "What are the pros and cons of human and automatic mapping and how can they complement each other?" By pointing out the difficulties in specific cases or groups of cases and grouping the sample into simple and difficult types of mappings, we show the limitations of current automatic methods and come up with some basic recommendations on what approach to use when.

Keywords: mapping thesauri; knowledge organization systems; intellectual mapping; ontology matching.

1. Introduction

Information on the Internet is constantly growing and with it the number of digital libraries, databases and information management systems. Each system uses different ways of describing their metadata, and different sets of keywords, thesauri and other knowledge organization systems (KOS) to describe its subject content. Accessing and synthesizing information by subject across distributed databases is a challenging task, and retrieving all information available on a specific subject in different information systems is nearly impossible. One of the reasons is the different vocabularies used for subject indexing. For example, one system might use the keyword 'snakes', whereas the other system uses the taxonomic name 'Serpentes' to classify information about the same subject. If users are not aware of the different 'languages' used by the systems, they might not be able to find all the relevant information. If, however, the system itself "knows", by means of mappings, that 'snakes' maps to 'Serpentes', the system can appropriately translate the user's query and therefore retrieve the relevant information without the user having to know about all synonyms or variants used in the different databases.

Mapping major thesauri and other knowledge organization systems in specific domains of interest can therefore greatly enhance the access to information in these domains. System developers for library search applications can programmatically incorporate mapping files into the search applications. The mappings can hence be utilized at query time to translate a user query into the terminology used in the different systems of the available mappings and seamlessly retrieve consolidated information from various databases[5].

Mappings are usually established by domain experts, but this is a very labor intensive, time consuming and error-prone task (Doerr, 2001). For this reason, great attention is being devoted to the possibility of creating mappings in an automatic or semi-automatic way (Vizine-Goetz, Hickey, Houghton, Thompsen, 2004), (Euzenat & Shvaiko, 2007), (Kalfoglou & Schorlemmer, 2003) and (Maedche, Motik, Silva, Volz, 2002). However, so far, research has focused mainly on the quantitative analysis of the automatically obtained mappings, i.e. purely in terms of precision and recall of either end-to-end document retrieval or of the quality of the sets of mappings produced by a system. Only little attention has been paid to a comparative study of manual and automatic mappings. A qualitative analysis is necessary to learn how and when automatic techniques are a suitable alternative to high-quality but very expensive manual mapping. This paper aims to fill that gap. We will elaborate on mappings between three KOS in the agricultural domain: AGROVOC, NALT and SWD.

- AGROVOC[6] is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). The AGROVOC Thesaurus was developed by the Food and Agriculture Organization of the United Nations (FAO) and the European Commission, in the early 1980s. It is currently available online in 17 languages (more are under development) and contains 28,718 descriptors and 10,928 non-descriptors in the English version.
- The NAL Thesaurus[7] (NALT) is a thesaurus developed by the National Agricultural Library (NAL) of the United States Department of Agriculture and was first released in 2002. It contains 42,326 descriptors and 25,985 non-descriptors organized into 17 subject categories and is currently available in two languages (English and Spanish). Its scope is very similar to that of AGROVOC. Some areas such as economical and social aspects of rural economies are described in more detail.
- The Schlagwortnormdatei[8] (SWD) is a subject authority file maintained by the German National Library and cooperating libraries. Its scope is that of a universal vocabulary. The SWD contains around 650,000 keywords and 160,000 relations between terms. The controlled terms cover all disciplines and are classified within 36 subject categories. The agricultural part of the SWD contains around 5,350 terms.

These controlled vocabularies (AGROVOC, NALT, and SWD) have been part of two mapping initiatives, conducted by the Ontology Alignment Evaluation Initiative (OAEI) and by the GESIS Social Science Information Centre (GESIS-IZ) in Bonn.

The **Ontology Alignment Evaluation Initiative (OAEI)** is an internationally-coordinated initiative to form consensus on the evaluation of ontology mapping methods. The goal of the OAEI is to help to improve the work on ontology mapping by organizing an annual comparative evaluation of ontology mapping systems on various tasks. In 2006 and 2007 there was a task that consisted in mapping the AGROVOC and NALT thesauri, called the "food task." A total of eight systems participated in this event. For this paper we consider the results of the five best performing systems that participated in the OAEI 2007 food task: Falcon-AO, RiMOM, X-SOM, DSSim and SCARLET. Details about this task, the data sets used and the results obtained can be found on the website of the food task[9]. The mapping relations that participants could use were the

[5] See the implementation of such an automatic translation service in the German social sciences portal Sowiport, available at http://www.sowiport.de.

[6] http://www.fao.org/aims/ag_intro.htm

[7] http://agclass.nal.usda.gov/agt.shtml

[8] http://www.d-nb.de/standardisierung/normdateien/swd.htm

[9] http://www.few.vu.nl/~wrvhage/oaei2007/food.html. Both the results and gold standard samples are available in RDF format.

SKOS Mapping Vocabulary relations exactMatch, broadMatch, and narrowMatch, because these correspond to the thesaurus constructs most people agree on: USE, BT and NT.

In 2004, the German Federal Ministry for Education and Research funded a major terminology mapping initiative called **Competence Center Modeling and Treatment of Semantic Heterogeneity**[10] at the GESIS-IZ, which published its conclusion at the end of 2007 (see project report in Mayr & Petras, 2008a, to be published). The goal of this initiative was to organize, create and manage mappings between major controlled vocabularies (thesauri, classification systems, subject heading lists), initially centred around the social sciences but quickly extending to other subject areas. To date, 25 controlled vocabularies from 11 disciplines have been intellectually (manually) connected with vocabulary sizes ranging from 1,000-17,000 terms per vocabulary. More than 513,000 relations were constructed in 64 crosswalks. All terminology-mapping data is made available for research purposes. We also plan on using the mappings for user assistance during initial search query formulation as well as for ranking of retrieval results (Mayr, Mutschke, Petras, 2008). The evaluation of the value added by mappings and the results of an information retrieval experiment using human generated terminology mappings is described in (Mayr & Petras, 2008b, to be published). The AGROVOC – SWD mapping was created within this initiative in 2007.

2. Related Work

Many thesauri, amongst which AGROVOC and the Aquatic Sciences and Fisheries Abstracts Thesaurus (ASFA) [11] are being converted into ontologies, in order to enhance their expressiveness and take advantage of the tools made available by the semantic web community. Therefore, great attention is being dedicated also to mapping ontologies. An example is the Networked Ontologies project (NeOn)[12], where mappings are one of the ways to connect ontologies in networks.

Within NeOn, an experiment was carried out to automatically find mappings between AGROVOC and ASFA. Since ASFA is a specialized thesaurus in the area of fisheries and aquaculture, the mapping with AGROVOC resulted in a mapping with the fisheries-related terms of AGROVOC. The mappings were extracted by means of the SCARLET system (cf. section 3) and were of three types: superclass/subclass, disjointness and equivalence. Evaluation was carried out manually by two FAO experts, in two runs: first with a sample of 200 randomly selected mappings, then with a second sample of 500 mappings. The experts were also supported in their evaluation by the graphical interface. The results obtained were rather poor (precision was 0.16 in the first run of the evaluation and 0.28 in the second run), especially if compared with the high results obtained by the same system with the mapping of AGROVOC and NALT (cf. section 3). The hypothesis formulated to explain this low performance is related to the low degree of overlap between AGROVOC and ASFA,[13] and that the terms in ASFA may not be well covered by the Semantic Web, as required by SCARLET. Cases like this clearly show how beneficial it would be to gain a clear understanding of when manual mapping is more advisable than automatic mapping (as in the case of the AGROVOC- ASFA mapping) or the other way around (as in the case of the AGROVOC - NALT mapping analyzed in this paper).

Another major mapping exercise was carried out mapping AGROVOC to the Chinese Agricultural Thesaurus (CAT) described in (Liang et al., 2006). The mapping has been carried out using the SKOS Mapping Vocabulary[14] (version 2004) and addresses another very important issue in mapping thesauri and other KOS: multilinguality. AGROVOC has been translated from

[10] The project was funded by BMBF, grant no. 01C5953. http://www.gesis.org/en/research/information_technology/komohe.htm.

[11] http://www4.fao.org/asfa/asfa.htm.

[12] http://neon-project.org.

[13] In particular, a problem could be the different level of details of the two resources, as ASFA tends to be very specific on fisheries related terms.

[14] http://www.w3.org/2004/02/skos/mapping/spec/.

English to Chinese, whereas CAT has been translated from Chinese to English. This creates potential problems as the following example illustrates: CAT '水稻'/'Oryza sativa' was originally mapped to AGROVOC 'Oryza sativa'. However, upon closer examination, the Chinese lexicalization in AGROVOC of 'Oryza sativa', which is '稻', appears to be the broader term of the CAT Chinese term. Moreover, a search in AGROVOC for the CAT Chinese term '水稻', shows the English translation as 'Paddy'. These discrepancies indicate the weakness of the above mentioned procedure and the necessity of cross checking all lexicalizations in both languages. Such cases pose hard problems for automatic mapping algorithms and can only be addressed with human support at the moment.

Other related work on semantic interoperability can be found in (Patel et al., 2005).

3. The AGROVOC – NALT Mapping within the OAEI

In the OAEI 2007 food task, five systems using distinct mapping techniques were compared on the basis of manual sample evaluation. Samples were drawn from each of the sets of mappings supplied by the systems to measure precision. Also, a number of small parts of the mapping were constructed manually to measure recall. Details about the procedure can be found in (Euzenat et al., 2007). Each participant documented their mapping method in a paper in the Ontology Matching 2007 workshop[15] (Hu et al., 2007), (Li, Zhong, Li, Tang, 2007), (Curino, Orsi, Tanca, 2007), (Nagy, Vargas-Vera, Motta, 2007) and (Sabou, Gracia, Angeletou, d'Aquin, Motta, 2007). Table 1 summarizes, for each system, the type of mapping found, how many mappings were identified and the precision and recall scores measured on the set of returned mappings, where:

Precision = | found mappings ∩ correct mappings | / | found mappings | and

Recall = | found mappings ∩ correct mappings | / | correct mappings |.

TABLE 2. The OEAI 2007 food task. Results (in terms of precision and recall) of the 5 systems participating in the initiative. Best scores are in **boldface**. All systems found equivalence mappings only, except SCARLET that also found hierarchical mappings.

System	Falcon-AO	RiMOM	X-SOM	DSSim	SCARLET		
Mapping type	=	=	=	=	=	< >	null(0)
# mappings	15300	18419	6583	14962	81	6038	647
Precision	**0.84**	0.62	0.45	0.49	0.66	0.25	
Recall	**0.49**	0.42	0.06	0.20	0.00	0.00	

The system that performed best at the OAEI 2007 food task was Falcon-AO. It found around 80% of all equivalence relations using lexical matching techniques. However, it was unable to find any hierarchical relations. Also, it did not find relations that required background knowledge to discover. This led to a recall score of around 50%. The SCARLET system was the only system that found hierarchical relations using the semantic web search engine Watson[16] (Sabou et al., 2007). Many of the mappings returned by SCARLET were objectively speaking valid, but more generic than any human would suggest. This led to a very low recall score.

4. The AGROVOC – SWD Mapping in the GESIS-IZ Approach

The GESIS-IZ approach considers intellectually (manually) created relations that determine equivalence, hierarchy (i.e. broader or narrower terms), and association mappings between terms from two controlled vocabularies. Typically, vocabularies will be related bilaterally, that means there is a mapping relating terms from vocabulary A (start terms in Table 2) to vocabulary B (end terms) as well as a mapping relating terms from vocabulary B to vocabulary A. Bilateral relations

[15] http://www.om2007.ontologymatching.org/
[16] http://watson.kmi.open.ac.uk/

are not necessarily symmetrical. E.g. the term ‘Computer’ in system A is mapped to term ‘Information System’ in system B, but the same term ‘Information System’ in system B is mapped to another term ‘Data base’ in system A. Bilateral mappings are only one approach to treat semantic heterogeneity; compare (Hellweg et al., 2001) and (Zeng & Chan, 2004). The approach allows the following 1:1 or 1:n mappings: Equivalence (=) means identity, synonym, quasi-synonym; Broader terms (<) from a narrower to a broad; Narrower terms (>) from a broad to a narrower; Association (^): mapping between related terms; and Null (0) which means that a term can not be mapped to another term. The first three of these relations correspond to the exactMatch, broadMatch, and narrowMatch relations from the SKOS Mapping Vocabulary.

The AGROVOC-SWD mapping is a fully human generated bilateral mapping that involves major parts of the vocabularies (see Table 2). Both vocabularies were analysed in terms of topical and syntactical overlap before the mapping started. All mappings in the GESIS-IZ approach are established by researchers, terminology experts, domain experts, and postgraduates. Essential for a successful mapping is the complete understanding of the meaning and semantics of the terms and the intensive use of the internal relations of the vocabularies concerned. This includes performing lots of simple syntactic checks of word stems but also semantic knowledge, i.e. to lookup synonyms and other related or associated terms.

TABLE 3. Mapping of Agrovoc – SWD. Numbers of established mappings by type and by direction.

Mapping direction	**# mappings**	**=**	**<**	**>**	**^**	**null 0**	**start terms**	**end terms**
AGROVOC - SWD	6254	5500 (4557 identical)	100	314	337	3	6119	6062
SWD - AGROVOC	11189	6462 (4454 identical)	3202	145	1188	192	10254	6171

The establishment of mappings is based on the following practical rules and guidelines:

1. During the mapping of the terms, all existing intra-thesaurus relations (including scope notes) have to be used.
2. The utility of the established relations has to be checked. This is especially important for combinations of terms (1:n relations).
3. 1:1 relations are prior.
4. Word groups and relevance adjustments have to be made consistently.

In the end the semantics of the mappings are reviewed by experts and samples are empirically tested for document recall and precision (classical information retrieval definition). Some examples of the rules in the KoMoHe approach can be found in (Mayr & Petras, 2008a, to be published).

5. Qualitative Assessment

Given these two approaches, one completely carried out by human subject experts and the other by machines trying to simulate the human task, the basic questions are: who performs more efficiently in a certain domain?, what are the differences?, and where are the limits? In order to draw some conclusions, a qualitative assessment is needed.

5.1. Alignment of the Mappings

We first “aligned” the mappings for the overlapping AGROVOC terms that have been mapped both to NALT and to SWD. For this we aligned the AGROVOC term with the mapped NALT terms (in English) and the mapped SWD term (in German): about 5,000 AGROVOC terms have been mapped in both approaches. For the AGROVOC-NALT mapping, we took the entire set of suggestions made by five systems participating in OAEI 2007. We also listed the number of systems that have suggested the mapping between the AGROVOC and the NALT term (between 1 and 5) and the specific mapping that has been assigned in the SWD mapping (equality, broader, narrower or related match). In case of several suggestions for a mapping (For example the AGROVOC term 'Energy value' has been suggested to be mapped to 'energy' as well as 'digestible protein' in the NAL thesaurus; the latter being an obvious mistake made by one of the systems.) we left all the multiple suggestions to be evaluated later.

We then grouped the aligned mappings into the higher level subject categories of AGROVOC and finally into four major terminology groups: Taxonomic, Biological/Chemical, Geographic, and Miscellaneous. These categories are the same as those used in the OAEI food task evaluation.

This was done in order to be able to draw more detailed conclusions on the difficulty of mappings based on the terminology group a particular mapping falls into. These groups were chosen in order to be more specific on whom to contact to evaluate the respective mappings. This will give an indication on what kind of knowledge is generally harder for automatic computer systems to map and what kind of background knowledge might also be needed to solve the more difficult cases.

5.2. Rating of a Representative Sample

Out of the about 5,000 mappings, we chose a representative sample of 644 mappings to be manually assessed. The mappings for the sample have been picked systematically in such a way that each of the groups is represented. We then assigned one of the following 6 difficulty ratings once for each of the mappings, AGROVOC-NALT and AGROVOC-SWD respectively. The assessments were done by Gudrun Johannsen and Willem Robert van Hage.

Table 3 summarizes our rating.

TABLE 4. Scale used to rate the mapping based on their "difficulty." The scale goes from 1 (Simple) to 6 (Hard Background Knowledge).

Rating	Explanation
1. Simple	the prefLabels are literally the same / exact match
2. Alt Label	there is a literal match with an alternative label / synonym in the other thesaurus
3. Easy Lexical	the labels are so close that any laymen can see that they are the same terms/concepts
4. Hard Lexical	the labels are very close, but one would have to know a little about the naming scheme used in the thesaurus (e.g. some medical phrases have a different meaning when the order of the words is changed and doctors know that)
5. Easy Background Knowledge	there are no clues as in point 1-4 for a match, but the average adult laymen knows enough to conclude that there is a mapping
6. Hard Background Knowledge	there are no clues as in point 1-4 for a match and you have to be an expert in some field, e.g. agriculture, chemistry, or medicine, to deduce that there is a mapping

5.3. Analysis of Examples

The assessment of the sample selection of 644 mappings is summarized in Table 4. The table is grouped by major subject groups: Taxonomic, Biological/Chemical and Miscellaneous. For each mapping approach (AGROVOC-NALT and AGROVOC-SWD), the table shows, what percentage of the mappings in the respective group are Simple, Easy Lexical, etc. The numbers in brackets are the absolute numbers. For example in the group Miscellaneous: 18.12% of the AGROVOC- SWD mappings in this subject group have been found to be of difficulty 6 (Hard Background Knowledge), whereas only 1.45% of the AGROVOC-NALT mappings have been given this rating.

Table 5 shows the mappings that have been wrongly assigned with the automatic approach in the AGROVOC-NALT mapping. In the assessment, we have specified if these wrong mappings should have been broader mappings (>), narrower mappings (<), related term mappings (^) or simply completely wrong, i.e. null (0) and should not have been suggested.

The Geographic group has been left out from the table, since the sample contained only very few mappings (less than 20). In any case, we can make the rather trivial statement that the Geographic group turns out to be rather simple, i.e. there seems to be an overall consensus on country names and other geographic concepts (in our case, the geographic group consists basically of country names). However, we have to be careful with this statement, especially when it comes to geopolitics. Borders of countries and similarly sensitive concepts might be called the same in two systems (and therefore seem simple and would be suggested by an automatic mapping tool with high security), but actually defined differently and mapping the two could raise sensitive issues. Take for example ‘Taiwan’: In AGROVOC, the non-preferred term ‘China (Taiwan)’ refers to the preferred term ‘Taiwan’, which has the broader term (BT) ‘China’, whereas in NALT ‘Taiwan’ uf ‘China (Taiwan)’ has the broader term ‘East Asia’. Another example, which is currently an issue, is the concept ‘Macedonia’. It has been used in the Codex Alimentarius[17] to refer to the former Yugoslavian Republic of Macedonia. However, since there is also a region in Greece, which is called Macedonia, the Greek authorities have requested the Codex Alimentarius to use ‘The former Yugoslavian Republic of’ in the name of the concept. Moreover, country definitions are very time dependent. How a user might best map geographical terms depends on the use case. For some purposes automatic mapping is a quick and good solution. For others it might be better to map all geographical terms manually, which is generally feasible due to the relatively small number of countries in the world (as compared, for example, to plant species).

TABLE 5. Rating of the mappings by terminology groups (taxonomic, biological, miscellaneous) and by rating of difficulty.

Taxonomic	Simple	Alt Label	Easy Lexical	Easy Background	Hard Lexical	Hard Background
AG - SWD	26.82%(70)	39.08%(102)	6.90%(18)	3.45%(9)	6.51%(17)	17.24%(45)
AG - NALT	65.13%(170)	22.61%(59)	1.15%(3)	0.00%(0)	1.92%(5)	0.00%(0)
Biological /Chemical	Simple	Alt Label	Easy Lexical	Easy Background	Hard Lexical	Hard Background
AG - SWD	62.35%(53)	21.18%(18)	1.18%(1)	2.35%(2)	1.18%(1)	11.76%(10)
AG - NALT	64.71%(55)	12.94%(11)	3.53%(3)	0.00%(0)	3.53%(3)	1.18%(1)
Miscellaneous	Simple	Alt Label	Easy Lexical	Easy Background	Hard Lexical	Hard Background
AG - SWD	33.33%(92)	11.96%(33)	10.14%(28)	16.67%(46)	9.78%(27)	18.12%(50)
AG - NALT	49.28%(136)	24.28%(67)	3.99%(11)	0.36%(1)	1.81%(5)	1.45%(4)

[17] The Codex Alimentarius Commission was created in 1963 by FAO and WHO to develop food standards, guidelines and related texts such as codes of practice under the Joint FAO/WHO Food Standards Programme. The main purposes of this Programme are protecting health of the consumers, ensuring fair trade practices in the food trade, and promoting coordination of all food standards work undertaken by international governmental and non-governmental organizations. It is available at: http://www.codexalimentarius.net/web/index_en.jsp.

TABLE 6. Mapping of AGROVOC-NALT. Classification of wrong mappings.

	should be <	should be >	should be null (0)	should be ^
Taxonomic	2.68%(7)	0.38%(1)	5.75%(15)	0.38%(1)
Biological / Chemical	2.35%(2)	1.18%(1)	10.59%(9)	0.00%(0)
Miscellaneous	1.45%(4)	0.36%(1)	13.77%(38)	3.26%(9)

Analyzing the other groups listed in the table leads to the few first statements: First of all, we can say that in general, Biological/Chemical like Geographical terminology is fairly easy to map (over 60% rated as Simple). This result makes sense, since like for geographical concepts there is probably a good consensus in the world on names of biological entities and chemicals[18]. Taking into account the alternative labels, this statement also holds for the group of taxonomic terminology mapping. Apparently, in the German language there are more discrepancies on the usage of preferred versus non-preferred labels and synonyms than in the English language. The Miscellaneous group (including the majority of mappings) appears to be the most difficult one: 13.77% of the automatically suggested mappings were even wrong, and it shows the highest percentage of Hard Background Knowledge mappings.

Further analyzing the mappings, we found that the AGROVOC-SWD mapping has a considerable amount of broader (>) and narrower (<) mappings. These are in general more difficult to find than equivalence mappings (either very easy or very difficult, because Hard Background Knowledge may be required), and therefore pose a big problem to automatic mapping algorithms. The SWD part on agriculture is also considerably smaller than the AGROVOC or NAL thesaurus and therefore many broader and narrower mappings are possible. Automatic mapping approaches have difficulty with such discrepancies. Apparently, subterms are often a good lexical clue for a < or > relation, but how does a computer decide which of the subterms is the superclass? Sometimes it is easy because one of the subterms is an adjective, while the other is a noun (e.g. 'mechanical damage' is a damage), but sometimes both are nouns (e.g. 'Bos taurus' is a Bos, not a taurus, but 'fruit harvester' is a harvester), and this is hard to parse. There are also cases where lexical inclusion can bring confusion, for example 'Meerrettich' (horseradish is *Armoracia rusticana*) and 'Meerrettichbaum' (horseradish tree is *Moringa oleifera*), as they refer to completely different concepts. Eventually, this problem might be solved by machine learning, but current mapping systems do not have any functionality to detect various common naming conventions.

It is remarkable that for the harder mappings (Hard Lexical, Easy Background, Hard Background), the percentage that has been found by the automatic approaches is overall very little (at most 3.53% for Hard Lexical biological/chemical terms), whereas the manual mapping approach can obviously identify these mappings. For example in the Miscellaneous group, more than 40% of the manual AGROVOC-SWD mappings fall into one of the three hardest ratings. The automatic mappings with this rating accumulate to less than 4%. Table 5 shows the numbers of wrong automatic mapping suggestions. The percentages in the three hardest ratings of the AGROVOC-NALT mapping are obviously cases of wrong suggestions, as listed in Table 5, which were either completely wrong mappings or should have been broader, narrower or related mappings.

It is not impossible, however, for automatic algorithms to also detect even Hard Background Knowledge mappings, for example by means of text mining. Some of these are easier to solve

[18] Organizations like The American Chemical Society (CAS, http://www.cas.org/expertise/cascontent/registry/) maintains lists of unique identifiers for chemicals in various languages. Various resources are also available that relate various chemical names to their CAS identifiers.

than others, because some background knowledge is simply easier to find. For instance, there are many web pages about taxonomy, but few about 'Lebensmittelanalyse' (food analysis). There are also many about chemicals, but few that state that a 'Heckstapler' (rear stapler) is some kind of 'Handhabungsgeraet' (handling equipment).

Some more concrete examples of mappings of varying difficulty:

1. *Mapping rated Alt label.* AGROVOC-NALT 'Marketing Strategies' = 'Marketing Techniques'. This mapping has been rated 'alt label', since, for example, in AGROVOC, 'Marketing Strategy' is the non-descriptor of 'Marketing Techniques'. This case makes it easy for an automatic classifier. However, this might also be misleading. In the agriculture domain, it might be correct to declare equivalence between these terms. However, in another domain there might actually be no mapping or at most a related term mapping. For example, in the business area, marketing strategies differ from marketing techniques substantially in that the strategies are long term objectives and roadmaps whereas the marketing techniques are operational techniques used in the marketing of certain products. For an automatic mapping algorithm, this is difficult to detect and alternative labels as they are sometimes found in thesauri, might be misleading.
2. *Mapping rated Hard Background Knowledge.* Both in AGROVOC and the NAL Thesaurus there is the term 'falcons' (exact match, simple mapping) while in SWD the German term 'Falke' does not exist, and thus had to be mapped to the broader term 'Greifvoegel' (predatory birds) which requires human background knowledge. However, in this case, the human knowledge could be found by a mapping system, if it would exploit the German Wikipedia. On the page about Falke[19], it states: "Die Falken (Gattung Falco) sind Greifvögel...".
3. *Mapping rated Hard Background Knowledge*. In SWD the term 'Laubfresser' (folivore) which does not exist in AGROVOC or in NALT had to be mapped to the broader term 'Herbivore'. This is another example where Hard Background Knowledge is needed.
4. Sometimes terms which seem to match exactly are incorrectly machine-mapped, for example when they are homonyms. Example: 'Viola' – in AGROVOC it is the taxonomic name of a plant (violets) while in SWD it refers to a musical instrument. In this case the relationship is 0. Sense disambiguation techniques such as the ontology partitioning performed by some of the current mapping systems, like Falcon-AO, should be able to solve most of these ambiguities by recognizing that none of the broader or narrower terms of 'Viola' and 'violet' are similar.

Some of the mappings of course will remain impossible for automatic methods that do not exploit sources of background knowledge, for example one of the AGROVOC-SWD mappings that found that 'Kater' (tomcat) is a 'männliches Individuum' (male individual).

6. Conclusion

The current mappings in the project at GESIS-IZ will be further analyzed and leveraged for distributed search not only in the sowiport portal but also in the German interdisciplinary science portal vascoda. Some of these mappings are already in use for the domain-specific track at the CLEF (Cross-Language Evaluation Forum) retrieval conference. We also plan on leveraging the mappings for vocabulary help in the initial query formulation process as well as for the ranking of retrieval results (Mayr, Mutschke & Petras, 2008).

We have seen that automatic mapping can definitely be very helpful and effective in case of Simple and Easy Lexical mappings. From our results, it appears that groups like Taxonomic vocabulary, Biological and Chemical Terminology and Geographic concepts fall into this category, as in general there seems to be more consensus on how to name things than in other groups. However, we need to be careful in these areas, where often word similarity does not mean

[19] http://de.wiktionary.org/wiki/Falke or http://de.wiktionary.org/wiki/Greifvogel.

that this is a potential mapping. These can be serious traps for automatic mapping approaches (like in the case of geopolitical issues).

Things get potentially more difficult in the case of more diversified groups/categories (in our case just summarized as Miscellaneous). Here, often background knowledge is needed to infer the correct mapping, and automatic mapping tools are able to identify only very little of these correctly. Most of the automatic suggestions are simply wrong or should not be equivalence relationships but broader, narrower or related terms.

The bottom line is that for the moment, mapping should not be seen as a monolithic exercise, but we can take the best of both approaches and use automatic mapping approaches to get to the simple and easy lexical mappings and then use human knowledge to control the ambiguous cases.

Acknowledgments

We would like to thank Lori Finch at the NAL for her extensive help on the AGROVOC-NALT mapping and for many discussions that contributed to this work. Van Hage was supported by the Dutch BSIK project Virtual Laboratory for e-science (http://www.vl-e.nl). The project at GESIS-IZ was funded by the German Federal Ministry for Education and Research, grant no. 01C5953. P. Mayr wishes to thank all our project partners and my colleagues in Bonn for their collaboration.

References

Curino, Carlo, Giorgio Orsi, and Letizia Tanca. (2007). X-SOM results for OAEI 2007. *Proceedings of the International Workshop on Ontology Matching.*

Doerr, Martin. (2001). Semantic problems of thesaurus mapping. *Journal of Digital Information*, *1*(8). Retrieved from http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/.

Euzenat, Jérôme, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Šváb, Vojtech Svátek, et al. (2007). *Results of the Ontology Alignment Evaluation Initiative*.

Euzenat, Jérôme, and Pavel Shvaiko. (2007). *Ontology matching*. Berlin, New York: Springer-Verlag.

Hellweg, Heiko, Jürgen Krause, Thomas Mandl, Jutta Marx, Matthias N. O. Müller, and Peter Mutschke, et al. (2001). *Treatment of semantic heterogeneity in information retrieval*. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 23). Retrieved from http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf.

Hu, Wei, Yuanyuan Zhao, Dan Li, Gong Cheng, Honghan Wu, and Yuzhong Qu. (2007). Falcon-AO: results for OAEI 2007. *Proceedings of the International Workshop on Ontology Matching*. Retrieved from http://www.dit.unitn.it/~p2p/OM-2007/5-o-Hu.OAEI.2007.pdf.

Kalfoglou, Yannis, and Marco Schorlemmer. (2003). Ontology Mapping: the state of the art. *Knowledge Engineering Review*, *18*(1), 1-31.

Li, Yi, Qian Zhong, Juanzi Li, and Jie Tang. (2007). Result of ontology alignment with RiMOM at OAEI'07. *Proceedings of the International Workshop on Ontology Matching.*

Liang, A., M. Sini, Chang Chun, Li Sijing, Lu Wenlin, He Chunpei and Johannes Keizer. (2006). The mapping schema from Chinese Agricultural Thesaurus to AGROVOC. *New review of hypermedia and multimedia*, *12*(1), 51-62. Retrieved from http://www.fao.org/docrep/008/af241e/af241e00.htm#Contents.

Maedche, Alexander, Boris Motik, Nuno Silva, and Raphael Volz. (2002). Mafra - a mapping framework for distributed ontologies. *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW).*

Mayr, Philipp, and Vivien Petras. (2008a to be published). Building a terminology network for search: The KoMoHe project. *Proceedings of the International Conference on Dublin Core and Metadata Applications.*

Mayr, Philipp, and Vivien Petras. (2008b to be published). Cross-concordances: Terminology mapping and its effectiveness for information retrieval. *Paper at the IFLA 2008, World Library and Information Congress: 74th IFLA General Conference and Council Québec, Canada*. Retrieved from http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf.

Mayr, Philipp, Peter Mutschke, and Vivien Petras. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. *Library Review*, *57*(3), 213-224.

Nagy, Miklos, Maria Vargas-Vera, and Enrico Motta. (2007). DSSim - managing uncertainty on the semantic web. *Proceedings of the International Workshop on Ontology Matching*.

Patel, Manjula, Traugott Koch, Martin Doerr, and Chrisa Tsinaraki. (2005). Semantic interoperability in digital library systems. Retrieved from http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/.

Sabou, Marta, Jorge Gracia, Sofia Angeletou, Mathieu d'Aquin, and Enrico Motta. (2007). Evaluating the semantic web: A task-based approach. *The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*.

Vizine-Goetz, Diane, Carol Hickey, Andrew Houghton, and Roger Thompsen. (2004). Vocabulary mapping for terminology services. *Journal of Digital Information, 4*(4). Retrieved from http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/.

Zeng, Marcia Lei, and Lois Mai Chan. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology, 55*(3), 377-395.

Full Papers

Session 3:

Metadata Generation: Methods, Profiles, and Models

Automatic Metadata Extraction from Museum Specimen Labels

P. Bryan Heidorn
University of Illinois
Champaign, IL, USA
pheidorn@uiuc.edu

Qin Wei
University of Illinois
Champaign, IL, USA
qinwei2@uiuc.edu

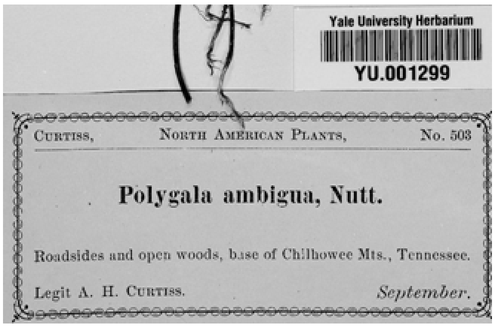
Abstract

This paper describes the information properties of museum specimen labels and machine learning tools to automatically extract Darwin Core (DwC) and other metadata from these labels processed through Optical Character Recognition (OCR). The DwC is a metadata profile describing the core set of access points for search and retrieval of natural history collections and observation databases. Using the HERBIS Learning System (HLS) we extract 74 independent elements from these labels. The automated text extraction tools are provided as a web service so that users can reference digital images of specimens and receive back an extended Darwin Core XML representation of the content of the label. This automated extraction task is made more difficult by the high variability of museum label formats, OCR errors and the open class nature of some elements. In this paper we introduce our overall system architecture, and variability robust solutions including, the application of Hidden Markov and Naïve Bayes machine learning models, data cleaning, use of field element identifiers, and specialist learning models. The techniques developed here could be adapted to any metadata extraction situation with noisy text and weakly ordered elements.

Keywords: automatic metadata extraction; machine learning; Hidden Markov Model; Naïve Bayes; Darwin Core.

1. Introduction

"Metadata can significantly improve resource discovery by helping search engines and people to discriminate relevant from non-relevant documents during an information retrieval operation" (Greenberg, 2006). Metadata extraction is especially important in huge and variable biodiversity collections and literature. Unlike many other sciences, in biology researchers routinely use literature and specimens going back several hundred years but finding the information resources is a major challenge. Metadata and data extracted from natural history museum specimens can be used to address some of the most important questions facing humanity in the 21st century including the largest mass extinction since the end of the age of the dinosaurs. What is the distribution of (the) species on earth? How has this distribution changed over time? What environmental conditions are needed by a species to survive?



Yale University Herbarium
YU.001299

CURTISS, NORTH AMERICAN PLANTS, No. 503

Polygala ambigua, Nutt.

Roadsides and open woods, base of Chilhowee Mts., Tennessee.

Legit A. H. CURTISS. *September.*

FIG. 1 Example Museum Specimen Label

There are over 1 billion specimens in museums worldwide collected over the past several hundred years. These specimens have labels (see Figure 1 for an example label) and catalog entries containing critical information including, the name of the species, the location and date of collection, revised nomenclature when the taxonomic name was changed, the habitat where it was found such as marsh or meadow as well as many other pieces of information. This knowledge will allow us to better predict the impact of global climate change on species distribution (Beaman, 2006). However, only a small fraction of this specimen data is available online. Consequently, digitization has become a high priority globally. Recent advances in digital imaging make it possible to quickly create images of specimen labels. However, the usefulness of the scanned images is limited since images cannot be easily manipulated and transformed into useful information in databases and full-text information systems. Optical Character Recognition (OCR) has proven to be useful but also challenging because of the age and variety of museum specimens. As is the situation with biomedical literature (Subramaniam, 2003), because of the volume and heterogeneity of the data it is difficult and expensive for humans to type in and extract critical information by hand. Automated and semi-automated procedures are required. “Results indicate that metadata experts are in favor of using automatic metadata generation, particularly for metadata that can be created accurately and efficiently. … metadata functionalities which participants strongly favored is running automatic algorithm(s) initially to acquire metadata that a human can evaluate and edit,” (Greenberg, 2006).

Research on museum labels is also important to other digitalization projects, eg. collection digitization in libraries. In general, the techniques developed here could be adapted to any information extraction situation of noisy text and with weakly ordered elements. In this paper, we discuss noisy-text extraction in more complex documents than in most prior works (e.g. Kahan,1987; Takasu, 2002; Takasu, 2003). Most noisy-text classification research is focused on how to automatically detect and correct the OCR errors, text segmentation, text categorization and text modeling (e.g. Takasu, 2002; Takasu, 2003; Foster, 2007). Some techniques that are used to reduce the effect of OCR introduced imperfections include: combining prior knowledge, N-grams, morphological analysis, and spatial information. Our research is focused on how to automatically extract metadata from noisy text using machine learning with limited training data. Since the output of handwriting OCR is still extremely poor, we limit our analysis below to labels that are primarily type written. Our experimental results demonstrate the effectiveness of exploiting tags within labels, and collection segmentation to improve performance.

The paper is organized as follows. Section 2 is a discussion of the properties of museum label metadata and information extraction challenges. Section 3, details how this problem has been addressed in other contexts, especially in the “address” and “bibliographic entry” problem. Section 4 details the system architecture, algorithm and the performance of the algorithms. Section 5 presents the conclusion and future work.

2. Metadata Properties

The research objective is to develop methods to extract an extended element set of Darwin Core (DwC) from herbarium records. DwC is an extensible data exchange standard for taxon occurrence data including specimens and observations of (once) living organisms. DwC has been adopted as a standard by the Biodiversity Informatics Standards (formerly the Taxonomic Database Working Group: http://darwincore.calacademy.org/). We extend the DwC to 74 fields that are particularly useful in museum specimen label context. Nearly 100% of the original label content can be assigned to some element. The 74 elements and their meanings are presented in Table 1. Some codes are optionally preceded with an “R” to indicate re-determination or appended with an “L” to indicate a field element label/identifier as discussed in section 4.3.

TABLE 1: 74 Elements and Element Meaning

Code	Element Meaning	Code	Element Meaning	Code	Element Meaning
ALT[L]	Altitude [Label]	HD	Header	PPREP	Possession Transfer Preposition
BC	Barcode	HDLC	Header Location	PTVERB	Possession Transfer Verb
BT	Barcode Text	[R]IN	[Re-determination] Institution	[R]SA	[Re-determination] Species Author
CD[L]	Collect Date [Label]	INLC	Institution Location	SC[L]	Species Code [Label]
CM[L]	Common Name [Label]	LATLON	Latitude and Longitude	[R]SN[L]	[Re-determination] Species Name [Label]
CN[L]	Collection Number [Label]	LC[L]	Location [Label]	SP	Species
CO[L]	Collector [Label]	MC[L]	Micro Citation [Label]	TC[L]	Town Code [Label]
CT	Citation	NS	Noise	TGN	Type Genus
DB[L]	Distributed By [Label]	OIN	Original Owning Institution	THD	Type Label Header
DDT[L]	Determination Date [Label]	OT	Other	TSA	Type Species Author
[R]DT[L]	[Re-]Determiner [Label]	PB[L]	Prepared By [Label]	TSP	Type Species
FM[L]	Family [Label]	PD[L]	Description [Label]	TY	Type Specimen
FT[L]	Footnote [Label]	PDT	Possession Transfer Date	[R]VAA	[Re-determination] Variety Author
[R]GN	[Re-determination] Genus	PIN	Possessing Institution	[R]VA[L]	[Re-determination] Variety [Label]
HB[L]	Habitat [Label]	PPERSON	Person Doing Possession Transfer		

The key problems with extracting information in this domain are heterogeneity of the label formats, open-ended vocabularies, OCR errors, and multiple languages. Collectors and museums have created label formats for hundreds of years so label elements can occur in almost any position and the any typography and script: hand written, typed and computer generated. In addition to typographic OCR errors, in these labels OCR error are also artifacts of format and misalignment (e. g. See ns(Noise) elements for OCR errors in the following xml example). These errors have several causes including: the later addition of data values to preprinted labels, label formats often included elements that are not horizontally aligned or because new labels were added to the original, making it difficult for OCR software to properly align the output. Following is the OCR output of the label in Figure 1 and the hand markup xml document. This markup is the target output for HLS and the format of the training and validation datasets. The tags indicate the semantic roles of the enclosed text.

OCR output of Figure 1:

```
^
¶£,&&¶
I ]    CUKTISS,
(}     -----------------
Poly gala ambigua, Nutt.
{¶>   Roadsides and open woods, b.ise of Chllhowec Mts., Tennessee.   5
Q
O   Legit, A. H. Cubtiss.
September.  9
```

XML markup of the OCRed text:

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="http://www3.isrl.uiuc.edu/~TeleNature/HERBIS/semanticrelax.rng" type="xml"?>
<labeldata><ns>^
 ¶£,&amp;&amp;¶
 I ]    </ns><co cc="Curtiss">CUKTISS,</co>
<ns> (}     -----------------
</ns><gn cc="Polygala"> Poly gala</gn><sp> ambigua,</sp><sa> Nutt.
</sa><ns> {¶&lt;</ns><hb>   Roadsides and open woods,</hb><lc cc="Base of Chilhowee Mts., Tennessee"> b.ise of Chllhowec Mts., Tennessee.</lc><ns>   5
 Q
 O</ns><col>   Legit,</col><co cc="A. H. Curtiss"> A. H. Cubtiss.
</co><cd> September.   9</cd>
 </labeldata>
```

3. Related Work

3.1. Evaluation Measures and Cross Validation

Before we introduce the related work, two important evaluation concepts are needed: F-score and K-fold cross-validation. The F-score is widely used in information retrieval and extraction evaluation and is calculated based on precision and recall (see following F equation). Generally speaking, the higher the F-score, the better the results. Precision is defined as the ratio of tokens correctly assigned to a class divided by the total number assigned to the class. The recall is the ratio of correctly classified tokens for a class divided by the total number of tokens of that class.

F = 2*precision*recall/(precision + recall) ***(F equation)***

Since trainings and validation data is expensive and time consuming to gather, K-fold cross-validation is frequently used to evaluate machine learning effectiveness in order to get more reliable results. Cross-validation is the statistical practice of partitioning a sample of training data into subsets so that machine learning is performed on one subset, while the other subset(s) are used to confirm the validity of the machine learning output (Witten, 2005). 5-fold, 10-fold and leave-one-out cross validation are very popular. In 5-fold validation the system randomly partitions the training set into five equal subsets. On each of 5 iterations, the machine learner will use one of the subsets for testing and use the other 4 sets as the training set.

3.2. Machine Learning Driven Information Extraction

Several automatic metadata extraction methods have been studied, e.g. hand-coded rule-based parsers (e.g. Han H. et al., 2005) and machine learning (e.g. Han, 2003; Borker, 2001). For highly structured tasks rule-based methods are easy to implement. The resulting rule system is usually domain-specific and can not be easily translated for use in other domains. Machine learning, on the other hand, is more robust and efficient (Han, 2003). Several learning models are available.

Among the most popular are the Naïve Bayes model (NB), the Hidden Markov Model (HMM), Support Vector Machines and Expectation Maximization. Supervised machine learning (SML) algorithms include training data and machine self-correction based on errors in machine performance against the training set. HMM and NB are discussed with more details in Section 4.

Substantial research has been conducted on the usefulness of ML in Information Extraction (IE). The most relevant prior research has been conducted on U.S. Postal address and bibliographic data. (e.g. Lewis, 1994; Frasconi, 2002; Borkar, 2001; Han, 2003; Hu, 2005, Cui, 2005). Borkar et al. developed a HMM system, similar to an algorithm we use, to handle the information extraction task (Borkar, 2001). The methods for "segmenting unformatted text records into structured elements" they proposed are successful in solving a simple U.S. postal address problem. They reported F-scores of 99%, 88.9% and 83.7% respectively on datasets of USPS addresses, Student addresses and Company addresses. For bibliographic data, they achieved an F-score of 87.3% for 205 records from Citeseer by using 100 training records. Han et al. implement a Support Vector Machine as the classifier to extract mainly Dublin Core metadata from Citeseer and EbizSearch, using 10-fold cross-validation on 500 training headers and 435 test headers. Their method achieves an overall accuracy of 92.9%. Cui's dissertation (2005) demonstrated that domain knowledge gained from machine learning models in one publication is very useful for improving the performance of IE in another publication in the same field. This is a necessary property of some machine learning algorithms we need to move the HERBIS Leaning System across herbarium collections.

Table 2 documents some of the differences between the address and the museum label information extraction problem and demonstrates the need for the new algorithms discussed below. This is an analysis of 200 U.S. addresses and 200 HERBIS (http://www.herbis.org/) "printed" label instances. The work cited above demonstrated that 200 records are sufficient for this type of analysis. The US address data and museum labels are randomly selected from regular USPS mail envelops and the HERBIS label database which includes more than 20,000 records from the Yale Peabody Herbarium. This herbarium was founded in 1864 and containing 350,000 specimens. We would expect similar results in similar Herbaria collections. Address labels and Museum labels were processed in exactly the same manner: image scanning followed by OCR and then markup. The museum label data is substantially more complex than the postal data (see Table 2).

TABLE 2. Statistics about experimental collections

Collection Statistics	**HERBIS**	**USPS Address**
Record count	200	200
Number of elements to recognize	74	10
Average number of words per instance	50	6.5
Approximate OCR error rate (error words/ total words)	15%	1%
Total number of element transitions	4736	969
Average fan-out factor	7.76	2.78
Average number of elements per instance	23.6	4.85
HMM F-score	76.9%	95.2%

The museum labels differ from prior datasets along a number of dimensions:

(1) Structure and order: Museum label data has a much looser structure than the address and bibliographic data. In spite of the fact that some of the museum labels are pre-printed and have a specific structure, there are still thousands of different formats. Some of the elements may appear anywhere of the original label (e.g. barcode, common name). Some elements are intertwined in natural language sentences. A particularly troublesome example is the mixing of habitat and location information e.g. "In boggy soil, 3.5 miles northeast of Deer Mountain."

The orderliness of these labels is reflected in the transitional probabilities. The transitional probabilities are the non-zero probability of one element follows another. This can be summarized in several ways, including the "Total Number of element transitions." This is a count of arcs connecting one element to another. This number is somewhat biased by the number of elements. The "Average Fan-out factor" is the average number of elements that can follow another. The value of 7.76 for museum labels means that on average any element can be followed by any 7 different elements.

(2) Variability within elements: Dictionary aided classification is usually unavailable in herbaria label data set. In Address problems, the proper name is an open class but the other elements are much more finite. The number of states, cities within a state and roads within a city are finite. In contrast, there are on the order of 1.5 million scientific names. The International Plant Name Index (IPNI) (http://www.ipni.org/) contains many thousands of entries but is far from complete particularly for older names that have been replaced yet appeared on museum labels. There are also variations in spelling because of human error or changes in nomenclatural rules. The list of all Collectors is also exceedingly long and labels do not follow any single authority. The location where a specimen was found is also an open class. It includes descriptions of locations, e.g. "300 meters NNW of the last rapids on Stanley Falls, Belgium Congo".

4. HERBIS Architecture

The museum domain is much more complex than the address problem as showed above and information extraction accuracy using the previously developed methods are inadequate. In this section we discuss methods we have used to enhance performance by extending the methods used for previous data sets. The goal of the learning phase of machine learning is to use representative examples to develop models that can, when presented with novel input, create proper classification of the input. Our first training data consists 200 digitized OCR records from the Yale Peabody Herbarium with multiple label formats randomly selected from the typed labels which requiring 10,095 element classifications.

4.1. Deployment

The HERBIS Learning System (HLS) is part of the overall HERBIS system. Museums anywhere in the world can create digital images of their specimens on their site. These images can be passed to the Yale Peabody Museum OCR processing unit where the label is detected and converted to a string sequence. This text packet is sent to HLS at UIUC though a web services connection. The text is converted to an XML document with appropriate information labeled and returns them to the end user. Other image handling services such as MorphBank (http://www.morphbank.net) can call the classification programs directly.

4.2. Learning Phrase: Application of HMM and Naïve Bayes

HLS uses a modified Hidden Markov Model(HMM). The HMM algorithm is discussed elsewhere (Borkar, 2001). The HMM induces a probability distribution on sequences of symbols. The HMM model is an order-preserving algorithm. There are three canonical problems associated with HMM could be solved by different algorithms. One of them is useful in information extraction context. Given the output sequence ($O_1\ O_2\ O_3 . . . O_t$), find the most likely sequence of hidden states ($S_1\ S_2\ S_3 . . . S_t$) that could have generated a given output sequence. In other words, given the word sequence ("Polygala ambigua, Nutt."), find the most likely sequences of element (i.e. gn(genus),sp(species),sa(species author)). This problem is solved by the Viterbi algorithm.

A Naïve Bayes (NB) model is a probability model based on conditional probabilities. The NB model makes predictions based on the probability distribution of features from the training set. The NB algorithm uses the distribution information to calculate the probabilities that a new instance belonging to the classes. The example would then be classified to the highest probability class. For computational efficiency NB assumes that each feature is conditionally independent of every other feature (Mitchell, 1997). This “independent” assumption greatly simplifies the model but the assumption is far from accurate in many cases. However, the overall classifier works surprisingly well in practice (Witten, 2005). The NB calculations are imbedded in part of the HMM algorithm.

In order to show the HMM performance comparing to others, we also implemented a non-ordered algorithm NB as the baseline and then present of series of extension to HMM below. The following example used the training data that was enhanced by including both the original OCR errors in the training set plus examples where the OCR errors were hand corrected. The difference from this correction is small so we only present the difference between HMM and NB on 41 elements that occur more than 20 times in the training set (Figure 2).

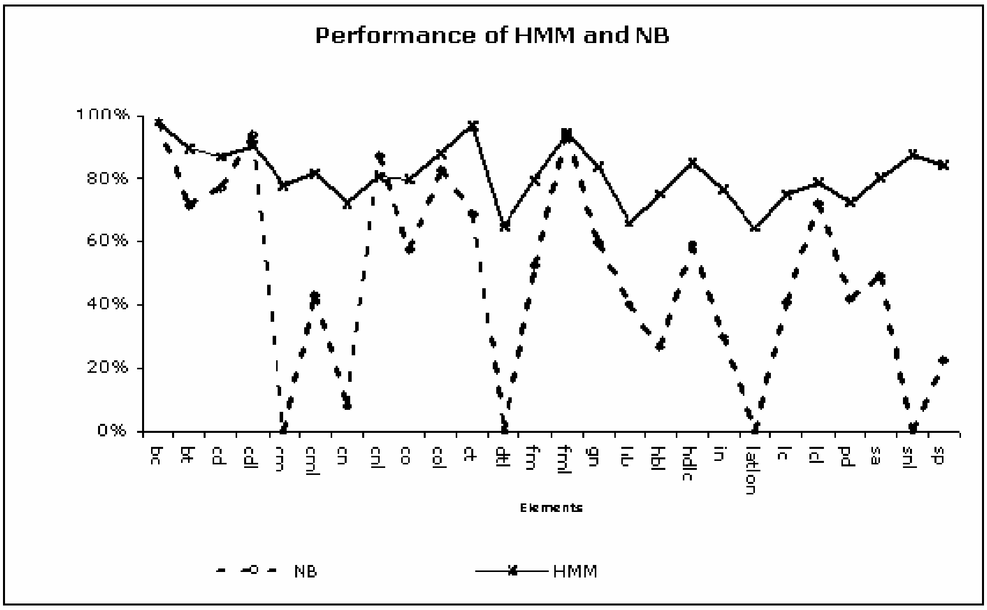


FIG. 2. Performance of HMM and NB

4.3. Field Element Identifiers

There is a set of elements in our dataset which we call “field element identifiers” (FEI). Some elements of some data labels are preceded by a string to identify the information that follows. For example, the term “Legit” in the string “Legit A. H. Curtiss” or “No.” in “No. 503” in Figure 1. In the museum label training data and machine learning output, we mark these with a terminal “L”, e.g. COL(collector label), LOL (location label), HBL (habitat label). Those label elements usually indicate that there is respectively a CO(collector), LC(location), HB(habitat) element following it, except in cases of missing data and alignment errors.

Rather than training the HMM algorithm to extract the Darwin Core elements and treat these other elements as NS(noise), we train the algorithms to recognize the field element identifiers as well. Our result shows that those label elements improve the ML overall 4%. Figure 3 presents the detailed performance differences between with label encoding in the schema and without those field element identifiers.

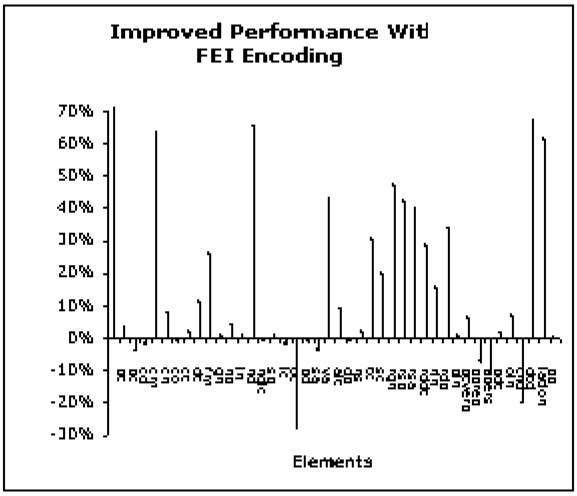


FIG. 3 Improved Performance With FEI Encoding

4.4. Dataset Segmentation and Social Computing (multiple User Feedback)

It is very difficult to improve the performance the ML without large numbers of training examples (Witten, 2005). Unfortunately, it is very expensive to get botanists to create these examples because creating the training examples from the raw OCR output is very time consuming.

The analysis above indicated that the performance difference between the USPS address and HERBIS collection are mainly attributable to the relative homogeneity in the format of the USPS addresses. There may be thousands of different formats of labels that have evolved over the last couple hundred years and now reside in museum collections. However, each collector has their own preferred format of label. This means that a particular museum will tend to have a relatively finite number of collectors supplying the museum at any one time and therefore will have a finite number of label formats represented in the collection. Further, if many museums are digitizing labels, then eventually, there will be corrected sets of labels for many collectors. It may be possible to develop multiple training modules each of which specializes in a particular collector and therefore label format. This observation leads to the hypotheses that the specialist model will perform better for records by the same author than for a generalized model trained on a random data collection and That fewer training examples will be required to reach a given level of performance using all labels from the same collector than would be required for a mixed collection of collectors. These hypotheses are supported in the results of the experiment below.

HLS includes the following Specialist Bootstrapping Architecture (SBA) (see Figure 4). Rather than following the standard machine learning model of creating training data >> generate model >> deploy model, we design a model where multiple museums could use available models to classify their data but as part of their workflow when they correct the machine learning data to put into their own database those examples are added to a new training pool. This pool can be subdivided into sub-collections to construct new specialist models (for particular collectors or collections).

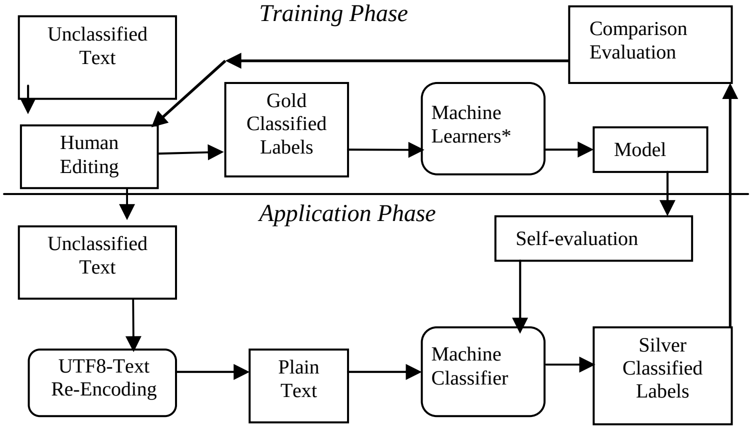


FIG. 4. Specialist Bootstrapping Architecture (SBA) for HERBIS
(*Machine Learners” in the diagram is one of many specialist learners.)

When the end-user sends a museum image to the server, the server would perform OCR, classify based on collector and then process the document with the appropriate collector or collection model. If a specialized collector module is not contained in the server, the information will be extracted from the label using the generic model based on a random sample of labels (see specialist learning algorithm below). For this strategy to work, it is necessary to be able to categorize labels into subsets prior to the information extraction step so that the highest performance model could be used for extraction. A Naïve Bayes pre-classifier can successfully perform this task. The 200 generic Yale training set includes 15 records from the collector “A. H. Curtiss”. The 5-fold evaluation of NB classifier trained to differentiate “Curtiss” from “non-Curtiss records” preformed well, F-Score of 97.5%.

Bootstrapping is a process where a small number of examples are used to create a weak learning model. This learning model, while weak is used to process a new set of examples. When a museum staff member corrects the output, it can be added to their database. The new result can help to form a stronger model. There are fewer errors generated by this new model making it easier for the users to correct the model’s errors. Museum staff who digitize records need to perform this step for key fields in any case in order to import the records to their database. These corrected examples are fed back into the process again to create an even stronger model. Successive generations of examples improve performance making it easier for the users to generate more examples.

A user wishing to create their own specialized model could begin by processing a set of labels from one collector through the generic Yale model. With each iteration the performance of the specialist system would improve but initially the generic model would perform better, with fewer errors per record. At some crossover point, the performance of the specialized model would exceed that of the generic model. In the example below the crossover point is at about 80 examples. In this framework the user only needs to correct machine output for 80 records to create a model that performs as well as a random collection of 200 records. This crossover point is what the algorithm is looking for in Phase 2 step 7 below.

Specialist Learning Algorithm --The steps could be described as follows:

Phase 1 (generic model)

1. Developers create a “generic” model alpha, M_0.
2. Developers create an empty training data set for User i (U_i) Training Set I, $\{T_i\}$.
3. Set best model $M_b = M_0$.
4. Go to Phase 2

Phase 2 (specialist model learning)

1. U_i runs a small unlabelled data set through M_b.
2. The system returns the newly labeled data (perhaps imperfect).
3. U_i fixes the errors, returns the fixed-labeled-data back to a learner.
4. The system adds the Records to $\{T_i\}$.
5. The system generates a new model M_i base on the $\{T_i\}$.
6. The system evaluates performance (p) of M_i and saves in performance log (L_i)
7. If $p(M_b) > p(M_i)$ set $M_b = M_i$
8. If U_i is satisfied with $p(M_i)$ got to Phase 3 else repeat Phase 2.

Phase 3 (specialized model application)

1 U_i runs any number of unlabelled data set through M_i.

2 The system returns the newly labeled data (perhaps imperfect).

4.5. Experiments and Result Analysis

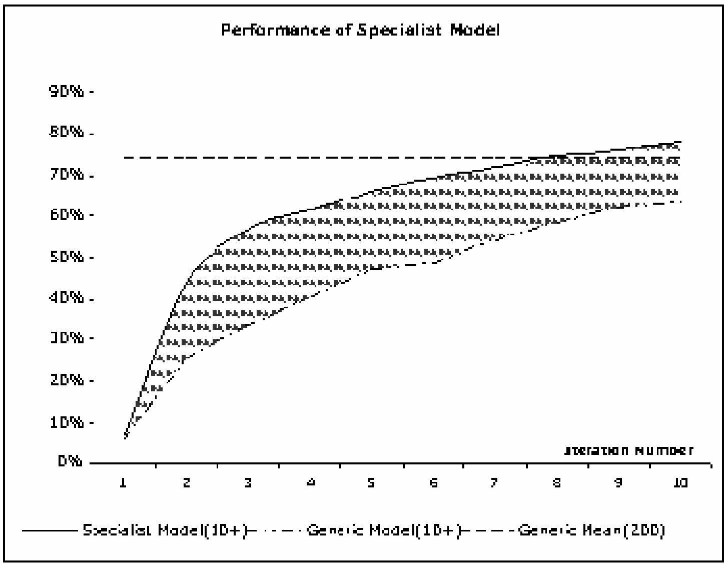


FIG. 5 Improved Performance of Specialist Model

This experiment compares the specialist model and the generic model generated from Yale 200 example collection. The dashed top line in figure 6 is the performance of 200 records independent of iteration. Regular expressions were applied to the 20,000 Yale digitized labels to identify the approximately 100 examples with the collector's name "A. H. Curtiss" who is a well-known collector and botanist. HLS was trained on 10 examples and then 5-fold evaluation used to measure the F-Score. This procedure was repeated 10 times, adding 10 new labels on each iteration producing a training set of 20, 30 and so on until a hundred were used in a training set. The results are presented in the solid curved line, "Specialist Model(10+)." Note that after the specialist model reaches 80 training records it matches the performance of the generic model trained on 200 randomly selected records. The dashed curved line at the bottom, Generic Model(10+), shows the performance of the learning algorithm when given comparable numbers of randomly selected training examples (not necessarily Curtiss) on each iteration. The shaded area is the advantage of using the specialist classification model. If we extended this dashed line out to 200 cases we would see the general model equal to the 200 case general Yale model. This is not demonstrated here since only 100 Curtiss examples exist in the 20,000 labels digitized at Yale. As predicted, fewer training examples are needed to reach a given level of performance using the Curtiss Specialist collection than a random collection. Given the effectiveness of the NB pre-classifier introduced in the previous section to identify collectors we should be able to create a specialist model for any collector. In fact, we can create a swarm of models for an arbitrary number of collectors and associated label types. The fact that there are only 100 Curtiss labels out of the 20,000 at Yale is a reflection of the fact that there are many labels and many formats.

5. Conclusion and Future work

Hidden Markov and Naïve Bayes models are potentially valuable tools for metadata extraction in herbarium labels but creation of sufficient data sets is a significant barrier to the application of machine learning. The number of required training examples and the associated work can be greatly reduced by establishing collaboration among museums digitizing their collections to support social machine learning. While the current system is a necessary prerequisite for an effective metadata generating system the machine learning swarm has not been implemented or tested with live data. Also, no sufficient user interface exists to deliver a functioning system. In creating such an interface a new set of research questions arise. Standard precision, recall and F-Scores are not sufficient for evaluating interactive systems. A more appropriate measure for botanists would be: How much time this system could save the expert when creating metadata? Important variables are the number of human corrections required per label, the time required to correctly complete a fixed number of labels, number of training examples and number of error corrections needed to meet some performance criteria such as a 90% F-score and other measures.

A number of options exist to improve underlying system performance. For example, label records might be processed in different orders to maximize learning and minimize error rate. OCR correction might be improved using context dependent automatic OCR correction. Dictionary lookup has been used extensively in automatic OCR correction. Context dependent correction means conducting the correct after knowing the word's class. For example, word "Ourtiss" should be corrected as "Curtiss". If the system already identified "Ourtiss" as collector, we can use the smaller collector dictionary instead of using a much larger general dictionary to do the correction. We proposed this method could get a better performance than just dictionary lookup.

Acknowledgements

This research was funded in part by the National Science Foundation, Grant #DBI-0345341.

References

Abney, Steven. (2002). Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA* (pp. 360-367).

Beaman, Reed S., Nico Cellinese, P. Bryan Heidorn, Youjun Guo, Ashley M. Green, and Barbara Thiers. (2006). HERBIS: Integrating digital imaging and label data capture for herbaria. *Botany 2006, Chico, CA.*

Borkar, Vinayak, Kaustubh Deshmuk, and Sunita Sarawagi. (2001). Automatic segmentation of text into structured records. *ACM SIGMOD, 30*(2), 175-186.

Cui, Hong. (2005). *Automating semantic markup of semi-structured text via an induced knowledge base: A case-study using floras*. Dissertation. University of Illinois at Urbana-Champaign.

Curran, James R. (2003). Blueprint for a high performance NLP Infrastructure. *Proceedings of the HLT-NAACL 2003 workshop on Software Engineering and Architecture of Language Technology Systems,* (pp. 39-44).

Foster, Jennifer. (2007). Treebanks gone bad: Parser evaluation and retraining using a Treebank of ungrammatical sentences. *IJDAR,* (pp. 129-145).

Frasconi, Paolo, Giovanni Soda, and Alessandro Vullo. (2002). Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, *18*(2-3), 195-217.

Greenberg, Jene, Kristina Spurgin, and Abe Crystal. (2006). Functionalities for automatic metadata generation applications: A survey of experts' opinions. *Int. J. Metadata, Semantics and Ontologies, 1*(1), 3-20.

Han, Hui, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhengyue Zhang, and Edward A. Fox. (2003). Automatic document metadata extraction using support vector machines. *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries,* (37–48).

Han, Hui, Eren Manavoglu, Hongyuan Zha, Kostas Tsioutsiouliklis, C. Lee Giles, and Xiangmin Zhang. (2005). Rule-based Word Clustering for Document Metadata Extraction. *ACM Symposium on Applied Computing 2005 March 13-17, 2005, Santa Fe, New Mexico, USA,* (pp. 1049-1053).

Hu, Yunhua, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, and Qinghua Zheng. (2005). Automatic extraction of titles from general documents using machine learning. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, June 07-11, 2005, Denver, CO, USA.*

Kahan S., Theo Pavlidis and Henry S. Baird. (1987). On the recognition of printed characters of any font and size. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 9*(2), 274-288.

Lewis, David D., and Marc Ringuette. (1994). A comparison of two learning algorithms for text categorization. *Proceedings of SDAIR, 3rd Annual Symposium on document Analysis and Information Retrieval*.

McCallum, Andrew K., and Dayne Freitag. (1999). Information extraction using HMMs and shrinkage. *Papers from the AAAI-99 workshop on Machine Learning for Information Extraction*.

Mehta, Rupesh, R. Pabitra Mitra, and Harish Karnick. (2005). Extracting semantic structure of web documents using content and visual information. *Special interest tracks and posters of the 14th international conference on World Wide Web WWW '05, Chiba, Japan,* (pp. 928-929).

Mitchell, Tom. M. (1997). *Machine learning*. McGraw Hill Higher Education.

Subramaniam, L. Venkata, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, et. al. (2003). Information extraction from biomedical literature: Methodology, evaluation and an application. *Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA,* (pp. 410-417).

Takasu, Atsuhiro and Kenro Aihara. (2002). DVHMM: Variable Length Text Recognition Error Model. *In Proceedings of International Conference on Pattern Recognition (ICPR02),* Vol.3, (pp. 110–114).

Takasu, Atsuhiro. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. *Proceedings of the 2003 Joint Conference on Digital Libraries.*

Witten, Ian H., and Eibe Frank. (2005). *Data mining: Practical machine learning tools and techniques (Second Edition).*

Achievement Standards Network (ASN): An Application Profile for Mapping K-12 Educational Resources to Achievement Standards

Stuart A. Sutton
University of Washington, USA
sasutton@u.washington.edu

Diny Golder
JES & Co, USA
dinyg@jesandco.org

Abstract

This paper describes metadata development of an application profile for the National Science Digital Library (NSDL) Achievement Standards Network (ASN) in the United States. The ASN is a national repository of machine-readable achievement standards modeled in RDF that shape teaching and learning in the various states. We describe the nature of the ASN metadata and the various uses to which that metadata is applied including the alignment of the standards of one state to those of another and the correlation of those standards to educational resources in support of resource discovery and retrieval.

Keywords: Resource Description Framework (RDF); educational resources; K-12 achievement standards; Achievement Standards Network (ASN); National Science Digital Library (NSDL)

1. Introduction

The correlation or mapping of learning resources such as lesson plans, curriculum units, and learning objects to formally promulgated achievement standards is a growing imperative in the U.S. K-12 environment. We choose "achievement standards" as a generic term indicating all forms of statements formally promulgated by a jurisdiction, community or organization to help shape teaching and learning in K-12 schools.[20] Achievement standards are frequently called curriculum objectives in the cataloging literature and academic standards, curriculum standards, learning indicators, benchmarks and an array of other names by various education communities and promulgating agencies. The standards movement in the U.S. has been stimulated largely by the perceived need to increase quality and accountability in the nation's K-12 schools.

> Starting slowly with the clarion call of *A Nation at Risk: The Imperative for Educational Reform*, development of policies defining accountability for U.S. teachers and schools has accelerated the processes of standards-based education in the U.S. Largely unheard of in the U.S. at the beginning of the 1990s, every state in the Union except one has promulgated achievement standards defining what K-12 students will learn, when that learning will take place, and how learning will be assessed. Influences such as the federal No Child Left Behind Act of 2001, testing regimes such as the National Assessment of Educational Progress (NAEP) and state high-stakes testing are major drivers in the developing call for learning resources that assist teachers in meeting the

[20] There are two broad classes of resources of concern to the ASN—curriculum standards and content standards. "[A] *content* standard describes what students should know and be able to do; a *curriculum* standard describes what should take place in the classroom. Specifically, curriculum standards address instructional technique or recommended activities as opposed to knowledge and skill per se (Marzano & Kendall, 1997)." "Content standards specify 'what students should know and be able to do.' They indicate the knowledge and skills—the ways of thinking, working, communicating, reasoning, and investigating, and the most important and enduring ideas, concepts, issues, dilemmas, and knowledge essential to the discipline—that should be taught and learned in school (National Education Goals Panel, 1993)"

> demands of demonstrable accountability lurking behind the articulated state standards (Sutton, 2008).[21]

The social and political thrust in the U.S. behind the national move toward accountability in K-12 education has roots in standards-based systems of teaching and learning. Since achievement standards reflect the knowledge, skills and habits of mind that K-12 students are expected to attain in a particular content area and at a given grade level, clear articulation of achievement standards coupled with rigorous assessment are at the heart of the systemic school initiatives in the U.S. under NCLB. NCLB invests states with the responsibility to create the standards for proficiency and then assess students against those standards in the core subjects of mathematics and language arts starting in 3rd grade. While the U.S. has what are loosely called "national standards," they are the result of standards-making activities of non-governmental organizations and bear no resemblance in terms of political force to the official national standards found in other countries around the world.

Fundamental to this notion of standards-based education are three "guiding questions" (Gaddy, Dean & Kendall, 2002):

1. What knowledge and skills will students be learning?
2. What evidence will be gathered and used to ensure that students learn?
3. What experiences will be used to ensure that students learn?

As illustrated in Figure 1, there should be a tight coupling among the achievement standards, what is being taught and the student learning assessed. Student learning degrades to the extent there is a misalignment between what is taught and what is assessed or between the goals of what is taught and the goals the educational system expects students to achieve.

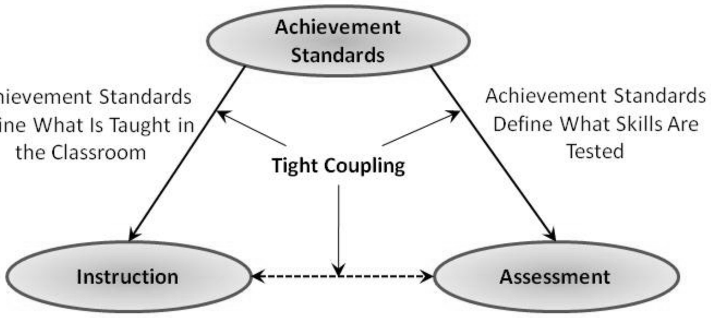

Achievement Standards
Achievement Standards Define What Is Taught in the Classroom
Achievement Standards Define What Skills Are Tested
Tight Coupling
Instruction
Assessment

FIG. 1: Aligning goals, content and assessment

One of the major goals of the ASN is to support this tight coupling amongst achievement standards, instruction and assessment by providing a national repository of comprehensive machine-addressable achievement standards that can be used by applications serving the education community including search engines, metadata generation tools and other 3rd party services. Prior to the ASN, collection holders and publishers wishing to correlate educational resources to achievement standards were faced with either developing very expensive, project-specific collections of achievement standards (and then maintaining them when they changed) or acquiring those standards from commercial entities (at even greater expense). In either case, the systems so deployed are not interoperable outside the closed system environments in which they were deployed.

The remainder of this paper is organized as follows: Section 2 describes the general architecture of the ASN in terms of its functionality and content. Section 3 is framed in terms of

[21] Since Sutton's paper was written in 2007, all 50 states in the U.S. have now created achievement standards for K-12 education.

the major tasks that the ASN is intended o accomplish in satisfying the need to express necessary relationships amongst the information objects of which the ASN is composed. Section 4 briefly explores the ASN potential for defining additional semantic relationships among ASN objects. In section 5, we describe the ASN mechanism for the refinement of standards statement by 3rd parties needing more fine-grained expressions than those provided by the promulgator in the canonical standards document. Section 6 provides conclusions and future directions.

2. ASN Architecture

The metadata for the ASN application profile has been developed around two primary objects—the K-12 standards document and the standards document component statements. The metadata for these objects, including declaration of relationships among them, has been modeled using Resource Description Framework (RDF). The modeling of the ASN took place during the early stages of the emergence of the DCMI Abstract Model and predates current DCMI work on the description set profile. Work is underway to bring the ASN XML/RDF encodings of the application profile into full alignment with the recent developments around the Abstract Model.

In order to guarantee maximum endorsement of the contents of the ASN by the promulgators of the standards, the focus of processing has been document-centric and the faithful rendering of the standards document in a form amenable to the Web environment. In document processing, each standards document is analyzed and decomposed into a set of atomic semantic units we call statements with each statement being assigned its own URI using Persistent URLs.

Several properties have been declared to express the structural relationships holding between individual statements and between statements and the parent document. It is anticipated that additional structural relationships among ASN objects will evolve as the publishing environment for standards matures and greater reliance is placed on the Web for access to those standards. In general, the current structural properties make it possible to express comprehensive units of meaning in standards documents in the form of hierarchical taxon paths. Figure 2 illustrates a single hierarchical taxon path for an Ohio math standard.

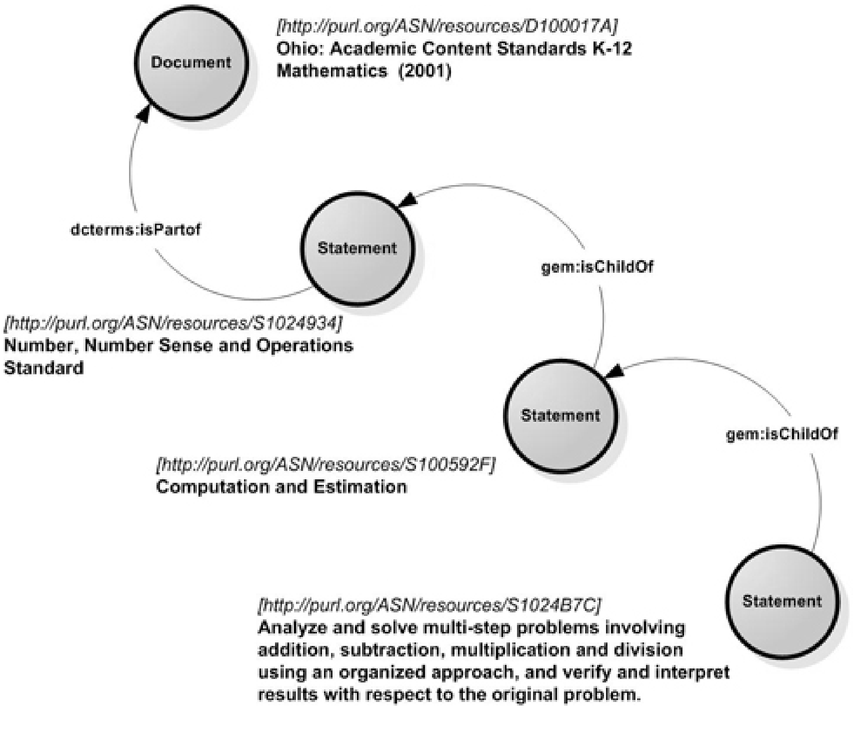

Document
[http://purl.org/ASN/resources/D100017A]
Ohio: Academic Content Standards K-12
Mathematics (2001)
dcterms:isPartof
Statement
gem:isChildOf
[http://purl.org/ASN/resources/S1024934]
Number, Number Sense and Operations
Standard
Statement
gem:isChildOf
[http://purl.org/ASN/resources/S100592F]
Computation and Estimation
Statement
[http://purl.org/ASN/resources/S1024B7C]
Analyze and solve multi-step problems involving
addition, subtraction, multiplication and division
using an organized approach, and verify and interpret
results with respect to the original problem.

FIG. 2. An Ohio math standard taxon path

The taxon path in Figure 2 is composed of metadata describing the standards document and a hierarchical structure of metadata describing three statements. Currently, two properties are used to handle these structural characteristics: (1) <dcterms:isPartOf> to describe the relationship between a statement and its standards document; and (2) <gem:isChildOf> to describe the hierarchical relationship between two statements in a taxon path. While the structure of most U.S. K-12 standards documents is hierarchical in nature, nothing in the ASN architecture precludes the definition of additional properties to manage more complex non-hierarchical structural relationships between statement objects.

Currently, access to the contents of the ASN repository of standards is accomplished either through: (1) the batch downloading of an entire standards document in RDF/XML from the ASN for use in local systems where the complete standards document is needed to meet local purposes; or, (2) through the dereferencing of an individual statement URI that has been assigned to a metadata record describing a resource. Dereferencing treats the object identified by the URI as the leaf object in a taxon path and returns all object metadata in the upward direction of the path including RDF/XML metadata describing the standards document.

2.1. Contents of the ASN

The ASN Achievement Content Standards Repository (ACSR) includes over 700 current and historical achievement standards documents for K-12 education as promulgated by departments or boards of education in each of the United States. Also included is a growing body of standards from nationally recognized content groups (e.g., the American Association for the Advancement of Science (AAAS)). Co-operative work is underway with the Australian Le@rning Federation to include all of the Australian national, state and territory standards. Currently, the machine-addressable standards statements in the ASN exceed 340,000 individual statements.

2.2. Functional Components of the ASN Architecture

The ASN architecture is composed of four major components and related services that make it possible for users and applications to access ACSR data stores and to author new standards documents within the ASN environment.

2.2.1. Standards Development Application (SDA)

The SDA assists standards bodies in developing well structured standards by providing a Web-based standards authoring environment. Created originally by ASN for the U.S. State Educational Technology Directors Association (SETDA), the development application is available to ASN member organizations maintaining standards within the ASN. Promulgators of achievement standards can register with the ASN and, through the authoring environment both author and publish their standards. The goal of the project with SETDA is for all 50 states in the U.S. to either author directly into the ASN or to republish in the ASN from their paper systems thus assigning globally unique ASN URI.

2.2.2. Standards Repository Application (SRA)

The SRA manages the ACSR data store of state, national, and international standards as well as the interface to the standards development application. The SRA also handles the processes associated with batch download of RDF/XML standards documents by third-party publishers, intermediaries and other service providers.

2.2.3. Metadata Generation Interface (MGI)

The MGI provides the means through Web Services for third-party metadata generation tools to interact with the ACSR and supports searching, browsing and the assignment of ASN URI to metadata records.

2.2.4. URI Resolver Application (URA)

The URA dereferences a state or national ASN standard URI embedded in a metadata record providing the full-text of each of the statement objects in the URI's associated taxon path.

3. Core ASN Tasks

As originally conceived, the ASN was intended to support two core tasks: (1) *correlation* of educational resources to achievement standards to support resource discovery and retrieval by K-12 teachers; and (2) *alignment* (mapping) of a standard statement in one standards document to a statement in a different standards document. Figure 3 illustrates these two core relationships.

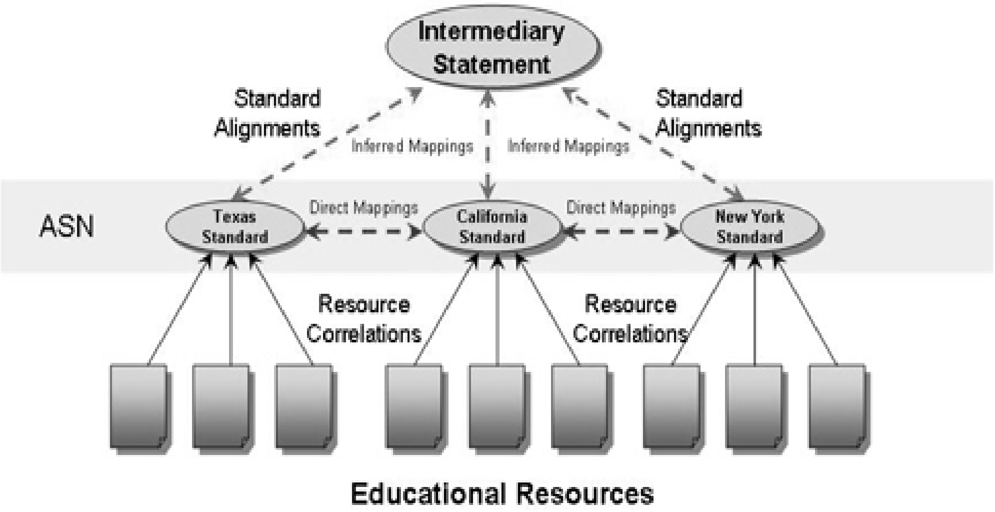


FIG. 3: Alignment and correlation processes

3.1. Correlation

A *correlation* is the assertion of a relationship between some educational resource and a standards statement as illustrated at the bottom of the Figure 3. In general, a correlation states that the resource being described is useful in achieving the goal(s) of the standards statement. In its simplest form, the <dcterms:conformsTo> property can be used in a Dublin Core description of an educational resource to assert this relationship. However, where the strength of fit between the resource being described and the standards statement is less than perfect, use of a separate description of the correlation including information regarding the strength of fit is more appropriate than the use of <dcterms:conformsTo>. We are in the process of defining a schema and accompanying constraints for describing such complex correlations in an educational resource description set where the educational resource being described is less than optimally useful in meeting the goal(s) of the standards statement.

3.2. Alignment

An *alignment* is the assertion of a relationship between a statement object in one standards document and a statement object in a different standards document—for example, the assertion that a statement in a Texas standard is similar to, or the same as, a statement in a New York standard. Thus, alignments are the means by which we make claims that one statement is more-or-less equivalent to another statement. Such alignments can be: (1) *direct* (see the red arrows in Figure 3), where the mappings are many-to-many; or (2) *inferred* where the mapping is to some form of intermediary statement and used in the manner of a switching language (i.e., many-to-one). The assumption behind the indirect alignments in Figure 3, is that we can state that there is a high likelihood that the substance of the Texas standard illustrated is similar to that of the New York standard because both are aligned to the same intermediary statement.

In the U.S., several NSF-funded research projects are developing intermediary applications that use ASN as their core data infrastructure. The Standards Alignment Tool (SAT) under development as part of the Computer-Assisted Content Standard Assignment & Alignment (CASAA) project at the Center for Natural Language Processing at Syracuse University uses natural language processing to suggest possible alignments between ASN standards statements (http://www.cnlp.org/research/project.asp?recid=48). WGBH uses ASN standards data in their Teachers Domain intermediary application that generates its alignment mappings dynamically through use of a controlled vocabulary performing the switching functioning (http://www.teachersdomain.org/). Through a member’s Teachers Domain profile, the system maps all retrieved educational resources to the controlling standards in the member’s state.

4. Semantic Relationships

While the current relationships defined in the ASN are structural in nature (e.g., defining the hierarchical structure of a taxon path as well as the structural relationship between a statement and its parent document), nothing in the ASN architecture precludes the definition of other semantic relationships between statement objects in one or more standards document objects. For example, strand maps, such as those developed by American Association for the Advancement of Science (AAAS) that incorporates the learning goals articulated in the *Benchmarks for Science Literacy* (Project 2061, 1993) and the Strand Map visualizations published in the *Atlas of Science Literacy* (Project 2061, 2001; Project 2061, 2007), help participants see how other standards statements relate and contribute meaning to the statement being studied. Thus, strand maps illustrate the relationships between individual learning goals and show the growth-of-understanding of ideas.

Edges connecting statements in the AAAS strand maps indicate that achieving the goal embodied in one statement contributes to achieving another. While the exact meaning of connecting lines in AAAS strand maps must be inferred from the context of the map, we envision making the meaning of various strand relationships explicit through definition of new ASN properties.

5. Refinement Semantics

Since ASN statements are faithful to the standards document, there are occasions when the granularity of a leaf in a taxon path could be effectively subdivided into more granular statements. For example, the leaf statement in the Ohio math standard from Figure 2 states:

> Analyze and solve multi-step problems involving addition, subtraction, multiplication and division using an organized approach, and verify and interpret results with respect to the original problem.

A publisher of testing instruments might well want to break this Ohio statement down into its sixteen constituent aspects (as illustrated in Table 1) in order to test separately one or more of those aspects.

TABLE 1: Statements derived from the canonical Ohio statement.

	involving addition	**involving subtraction**	**involving multiplication**	**involving division**
Analyze multistep problems …	analyze addition	analyze subtraction	analyze multiplication	analyze division
Solve multistep problems …	solve addition	solve subtraction	solve multiplication	solve division
Verify multistep problems …	verify addition	verify subtraction	verify multiplication	verify division
Interpret multistep problems …	interpret addition	interpret subtraction	interpret multiplication	interpret division

To accommodate the need to further refine what we call *original* statements (i.e., the canonical statement from the standard's promulgator), we define a class of *derived* statements. This process of refinement is illustrated in Figure 4.

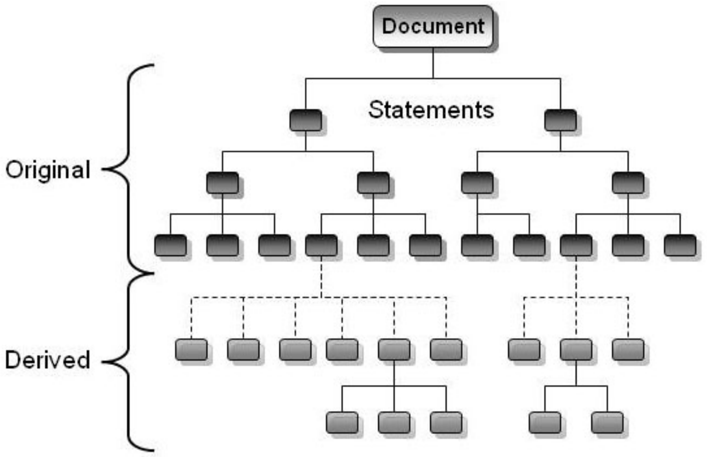


FIG. 4: Refining taxon paths through creation of derived statement objects by 3rd parties

In general, derived statement objects will be created either directly in the ASN by 3rd parties with the need for such refinements or in a namespace maintained by those parties. However, nothing precludes the ASN from creating such refinements where it deems it necessary to do so. In either case using the example in Figure 4, the derived statements are declared as children of the original statement created by the promulgating agency.

6. Conclusion & Future Work

The ASN is intended to provide critical system and data infrastructure to support K-12 teaching and learning in the U.S. It provides a common reference for any information system needing to utilize achievement standards in delivering interoperable standards-based services to the educational community. However, the ASN provides more than authoritative achievement standards texts in digital form by articulating a principled framework for future development of standards-based services that are amenable to the Semantic Web.

While any promulgating standards body can use the ASN to author and expose their standards, we are aware that other standards repositories will likely be developed—perhaps by the individual promulgators of some standards. What the work with the ASN provides is a means by which such systems can be designed to interoperate intelligently. System criteria for such interoperability include:

- Standards documents and their distinct semantic units (i.e., analogs of ASN statements) are treated as related objects within the system;
- Standards documents and each semantic unit are described (including the source text of each semantic unit);
- Each object in the system is assigned a URI that is dereferenceable by humans and Web-based applications; and
- The value returned through dereferencing is the set of URIs of the objects that compose the complete taxon path—thus providing everything necessary to reconstruct the structural and semantic context of the identified standard object.

Future work, in addition to the development of the separate correlation resource discussed briefly in Section 3.1, includes the exploration of versioning demands and mechanisms for standards statement objects. The document-centric nature of the ASN reflects the reality of the current publishing environment for U.S. K-12 standards. Promulgators of these standards periodically publish new versions with each version superseding the previous one. However, we think that as the publishing environment shifts to the Web through applications such as the ASN, fewer promulgators of standards will follow this publication cycle and will instead engage in ongoing versioning at the level of what the ASN defines as the statement or the taxon path.

In anticipation of such a shift, we are currently exploring mechanisms of statement versioning that track the changes to statement objects over time. This will allow us to aggregate 'families' of statement objects while maintaining metadata about each object-e.g., when a particular version of a statement object was created, under what circumstances, and how that object relates to other versions of the same statement.

The ASN work described here is somewhat related to the work of the IEEE LTSC 1484.20 Reuseable Competency Definitions (RCD) standard.[22] However, it differs in substantial ways including reliance on a different underlying abstract model and the RCD's focus on unstructured text intended for human interpretation. As work on the ASN goes forward and the Joint DCMI/IEEE LTSC Taskforce's work on expressing IEEE LOM metadata using the Dublin Core Abstract Model moves to completion, we anticipate that aspects of the RCD may be deployed in the ASN framework.[23]

Acknowledgements

The authors acknowledge the support of the entire ASN community including Bruce Walker and our colleagues at the Syracuse University Center for Natural Language Processing (CNLP), WGBH: Teachers Domain, DLESE and Digital Learning Sciences, the American Association for the Advancement of Science (AAAS), and the State Educational Technology Directors Association (SETDA). This work is partially funded by the National Science Foundation under grant number DUE: 0121717.

[22] IEEE LTSC 1484.20 Working Group, http://ltsc.ieee.org/wg20/.

[23] Joint DCMI/IEEE LTSC Taskforce, http://dublincore.org/educationwiki/DCMIIEEELTSCTaskforce.

References

Commission on Excellence in Education (NCEE). (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

Gaddy, Barbara B., Ceri B. Dean, and John S. Kendall. (2002). *Noteworthy perspectives: Keeping the focus on learning.* Aurora, CO: Mid-continent Research for Education and Learning. Retrieved March 21, 2007, from http://www.mcrel.org/PDF/Noteworthy/5022IR_NW_Focus.pdf.

Marzano, Robert J., and John S. Kendall. (1997). *The fall and rise of standards-based education: A National Association of School Boards of Education (NASBE) issues in brief.* Aurora, CO: Mid-continent Research for Education and Learning. Retrieved March 23, 2008, from http://www.mcrel.org/PDF/Standards/ 5962IR_FallAndRise.pdf.

National Education Goals Panel. (1993). *Promises to keep: Creating high standards for American students.* Washington, DC: National Education Goals Panel.

No Child Left Behind Act of 2001 (§§ 6301-7941). 20 U.S.C.

Project 2061, American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy.* New York: Oxford University Press.

Project 2061, American Association for the Advancement of Science. (2001). *Atlas of Science Literacy*. Washington, D.C.: American Association for the Advancement of Science and the National Science Teachers Association.

Project 2061, American Association for the Advancement of Science. (2007). *Atlas of Science Literacy*. Washington, D.C.: American Association for the Advancement of Science and the National Science Teachers Association.

Sutton, Stuart A. (2003). Principled design of metadata-generation tools for educational resources. In Marcia A. Mardis, *Developing digital libraries for K-12 education,* (pp. 45-63). Syracuse, NY: ERIC Clearinghouse on Information and Technology.

Sutton, Stuart A. (2008). Metadata quality, utility and the semantic web: The case of learning resources and achievement standards. *Cataloging & Classification, 46*(1).

Appendix A. ASN Taxon Path RDF/XML Encoding

```
<rdf:RDF
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:gemq="http://purl.org/gem/qualifiers/"
    xmlns:asn="http://purl.org/ASN/schema/core/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

  <!—STATEMENT DESCRIPTION -->
  <rdf:Description rdf:about="http://purl.org/ASN/resources/S1024934">
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/4"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/10"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/K"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/6"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/8"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/12"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/2"/>
    <dcterms:isPartOf rdf:resource="http://purl.org/ASN/resources/D100017A"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/9"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/11"/>
    <dcterms:description>Number, Number Sense and Operations Standard</dcterms:description>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/7"/>
    <rdf:type rdf:resource="http://purl.org/ASN/schema/core/Statement"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/1"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/3"/>
    <dcterms:subject rdf:resource="http://purl.org/ASN/scheme/ASNTopic/math"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/5"/>
  </rdf:Description>
  <!—STATEMENT DESCRIPTION -->
  <rdf:Description rdf:about="http://purl.org/ASN/resources/S100592F">
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/6"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/10"/>
    <dcterms:subject rdf:resource="http://purl.org/ASN/scheme/ASNTopic/math"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/1"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/12"/>
    <rdf:type rdf:resource="http://purl.org/ASN/schema/core/Statement"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/3"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/9"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/7"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/5"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/11"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/2"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/8"/>
    <dcterms:description>Computation and Estimation</dcterms:description>
    <gemq:isChildOf rdf:resource="http://purl.org/ASN/resources/S1024934"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/K"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/4"/>
  </rdf:Description>

  <!—DOCUMENT DESCRIPTION -->
  <rdf:Description rdf:about="http://purl.org/ASN/resources/D100017A">
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/6"/>
    <dcterms:created>2001</dcterms:created>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/10"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/8"/>
    <rdf:type rdf:resource="http://purl.org/ASN/schema/core/StandardDocument"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/12"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/7"/>
    <dc:title>Academic Content Standards K-12 Mathematics</dc:title>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/5"/>
    <dcterms:description xml:lang="en-US">The mathematics academic content standards prepare all students for
success in the workplace and post-secondary education. Competency in mathematics includes understanding of
mathematical concepts, facility with mathematical skills, and application of concepts and skills to problem-solving
situations. Students are able to communicate mathematical reasoning using mathematical
and everyday language.</dcterms:description>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/9"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/K"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/1"/>
    <asn:jurisdiction rdf:resource="http://purl.org/ASN/scheme/ASNJurisdiction/OH"/>
    <dcterms:subject rdf:resource="http://purl.org/ASN/scheme/ASNTopic/math"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/2"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/11"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/3"/>
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/4"/>
  </rdf:Description>

 <!—STATEMENT DESCRIPTION -->
  <rdf:Description rdf:about="http://purl.org/ASN/resources/S1024B7C">
    <dcterms:educationLevel rdf:resource="http://purl.org/ASN/scheme/ASNEducationLevel/4"/>
    <rdf:type rdf:resource="http://purl.org/ASN/schema/core/Statement"/>
    <dcterms:description>12. Analyze and solve multi-step problems involving addition, subtraction, multiplication
and division using an organized approach, and verify and interpret results with respect to the original
problem.</dcterms:description>
    <dcterms:subject rdf:resource="http://purl.org/ASN/scheme/ASNTopic/math"/>
    <gemq:isChildOf rdf:resource="http://purl.org/ASN/resources/S100592F"/>
  </rdf:Description>

</rdf:RDF>
```

APPENDIX B.

ASN Document Properties

Property Label	Context-Specific Definition
Adoption Date	The date the standards document was adopted by the jurisdiction in which it was intended to apply.
Creator	The person or organization chiefly responsible for the intellectual content of the standards document.
Change Note	A change note is intended for documenting fine-grained changes to a standards document for the purposes of administration and management.
Date Copyrighted	Date of the copyright of the standards document.
Date Valid	Date (often a range) of validity of a standards document.
Description	An account of the content of the standards document.
Editorial Note	Information regarding the analysis of the standards document in preparation for its representation within the ASN.
Education Level	The grade or grade bands covered by the standards document being described.
Has Child	Identifies child statements of the standards document being described. I.e., identifies the top-level statements of the standards document.
History Note	A piece of information intended for users of the scheme, documenting significant changes to the meaning/form/state of the standards document since any previous version.
Identifier	An unambiguous reference to the resource within a given context. For the ASN standards documents, the value of the «identifier» is always a network-resolvable URI.
Jurisdiction	A legal, quasi-legal, organizational or institutional domain of the entity mandating the use of the achievement standard--e.g., California.
License	A legal document giving official permission to do something with the standards document.
Local Subject	The text string denoting the subject of the document as designated by the promulgating agency.
Publisher	An entity responsible for making the resource available. In the ASN, the promulgating agency of the standards document.
Repository Date	The date the standards document was added to the ASN repository.
Rights	Information about rights held in and over the standards document.
Status	The publication status of the standards document--e.g., "Draft," "Published," "Superseded."
Subject	The ASN topic of the content of the document being described.
Title	A name given to the standards document by the promulgating agency.

ASN Statement Properties

Property Label	Context-Specific Definition
Creator*	A person or organization chiefly responsible for the intellectual content of the statement being described when different from the creator of the standards document (e.g., 3rd party derived statement).
Comment	Supplemental text provided by the promulgating body that clarifies the nature, scope or use of the statement being described.
Concept Term	A word or phrase used by the promulgating agency to refine and differentiate the statement being described contextually (e.g., a McREL concept term).
Created*	Date of creation of the statement.
Description	The text of the statement being described.
Education Level	The grade or grade bands covered by the standards statement being described.
Identifier	An unambiguous reference to the resource within a given context. For the ASN standards statement, the value of the «identifier» is always a network-resolvable URI.
Is Child Of	The statement being described is lower in some arbitrary hierarchy than the statement identified in the «isChildOf» property. The statement identified is a parent of the statement being described.
Is Part Of	The described statement is a physical or logical part of the referenced standards document.
Jurisdiction*	A legal, quasi-legal, organizational or institutional domain of the entity mandating the use of the statement--e.g., California.
Local Subject*	The text string denoting the subject of the statement as designated by the promulgating agency.
Relation*	A related resource.
Statement Label	The textual label identifying the class of the statement as designated by the promulgating body—e.g., "Standard," "Benchmark," "Strand," or "Topic."
Statement Notation	An alphanumeric notation or ID code as defined by the promulgating body to identify the statement.
Status	The publication status of the statement taken from the ASN Status controlled vocabulary.
Subject	An ASN topic of the content of the statement being described.

*Properties that are generally optional with ASN statements but mandatory when the statement is “derived” (i.e., created a 3rd party).

Collection/Item Metadata Relationships

Allen H. Renear
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship
University of Illinois, USA
renear@illinois.edu

Karen M. Wickett
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship
University of Illinois, USA
wickett2@illinois.edu

Richard J. Urban
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship
University of Illinois, USA
rjurban@illinois.edu

David Dubin
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship
University of Illinois, USA
ddubin@illinois.edu

Sarah L. Shreeves
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship
University of Illinois, USA
sshreeve@illinois.edu

Abstract

Contemporary retrieval systems, which search across collections, usually ignore collection-level metadata. Alternative approaches, exploiting collection-level information, will require an understanding of the various kinds of relationships that can obtain between collection-level and item-level metadata. This paper outlines the problem and describes a project that is developing a logic-based framework for classifying collection/item metadata relationships. This framework will support (i) metadata specification developers defining metadata elements, (ii) metadata creators describing objects, and (iii) system designers implementing systems that take advantage of collection-level metadata. We present three examples of collection/item metadata relationship categories, *attribute/value-propagation*, *value-propagation*, and *value-constraint* and show that even in these simple cases a precise formulation requires modal notions in addition to first-order logic. These formulations are related to recent work in information retrieval and ontology evaluation.

Keywords: metadata; Dublin Core; collections; context; logic; inferencing

1. Introduction

Collections of texts, images, artifacts, and other cultural objects are often designed to support specific research and scholarly activities. Toward that end collections themselves are carefully developed and described. These collection descriptions indicate such things as the purpose of the collection, its subject, the method of selection, size, nature of contents, coverage, completeness, representativeness, and a wide range of summary characteristics, such as statistical features. This information enables collections to function not just as aggregates of individual data items but as independent entities that are in some sense more than the sum of their parts, as intended by their creators and curators (Curral, Moss & Stuart, 2005; Heaney, 2000; Lagoze, et al. 2006 Lee, 2000, 2005; Palmer, 2004, 2006). Collection-level metadata, which represents this information in computer processable form, is thus critical to the distinctive intellectual and cultural role of collections as something more than a set of individual objects.

Unfortunately, collection-level metadata is often unavailable or ignored by contemporary retrieval and browsing systems, with a corresponding loss in the ability of users to find, understand, and use items in collections (Foulonneau, et al., 2005; Wendler, 2004). Preventing this loss of information is particularly difficult, and particularly important, for "metasearch", where item-level descriptions are retrieved from a number of different collections simultaneously, as is the case in the increasingly distributed search environment of the Internet (Christenson & Tennant, 2005; Dempsey, 2005; DLF, 2005; Foulonneau, et al., 2005; Lagoze, et al., 2006; Warner, et al., 2007).

The now familiar example of this challenge is the "on a horse" problem, where a collection with the collection-level subject "Theodore Roosevelt" has a photograph with the item-level annotation "on a horse" (Wendler, 2004). Item-level access across multiple collections (as provided not only by popular Internet search engines, but also specialized metasearch and federating systems, such as OAI portals) will not allow the user to effectively use a query with keywords "Roosevelt" and "horse" to find this item, or, if the item is retrieved using item-level metadata alone, to then use collection-level information to identify the person on the horse as Roosevelt.

The problem is more complicated and consequential than the example suggests and the lack of a systematic understanding of the logical relationships between collection-level metadata and item-level metadata is an obstacle to the development of remedies. This understanding is what is required not only to guide the development of context-aware search and exploitation, but to support curation policies as well.

The problem is also urgent: even as recent research confirms the key role that collection context plays in the scholarly use of information resources (Brockman, et al., 2001; Palmer, 2004), the Internet has made the context-free searching of multiple collections routine.

We are developing a framework for classifying and formalizing collection/item metadata relationships and determining inference rules that can be incorporated into retrieval and browsing systems. This undertaking is part of a larger project, recently funded by U.S. Institute for Museum and Library Services (IMLS), to develop tools for improved retrieval and exploitation across multiple collections.[24]

2. The DCC/CIMR Project

These issues were initially raised during an IMLS Digital Collections and Content (DCC) project, begun at the UIUC in 2003. That project developed a collection-level metadata schema based on the RSLP and Dublin Core Metadata Initiative (DCMI) and created a collection registry for all digital collections funded through the IMLS National Leadership Grant (NLG) since 1998, with some Library Services and Technology Act (LSTA) funded collections included. The registry currently contains records for 200 collections. An item-level metadata repository was also developed, which has harvested 76 collections using the OAI-PMH protocol. Our research initially focused on overcoming the technical challenges of aggregating large heterogeneous collections of item-level records and collection descriptions. We conducted studies on how content contributors conceived of the roles of collection descriptions in digital environments (Palmer & Knutson, 2004; Palmer et al., 2006), and preliminary usability work. These studies and related work on the CIC Metadata Portal[25], suggest that while the boundaries around digital collections are often blurry, many features of collections are important for helping users navigate and exploit large federated repositories, and that collection and item-level descriptions should work in concert to benefit certain kinds of user queries (Foulonneau, et al., 2005).

Concurrently, we studied the quality of the harvested item-level metadata using a range of qualitative and quantitative metrics. While the obstacles to building effective aggregations of item-level metadata are well documented (Arms et al., 2003; Dushay and Hillmann, 2003; Hutt

[24] IMLS Digital Collections and Content. http://imlsdcc.grainger.uiuc.edu/about.asp

[25] http://cicharvest.grainger.uiuc.edu/

and Riley, 2005), we were interested the quality dimensions that could be measured in order to better understand where poor quality might impede interoperability. Using an information quality framework proposed by Gasser and Stvilia (Gasser and Stvilia 2001; Stvilia et al. 2004) we found that the relational or contextual information quality dimensions—that is, the dimensions that depend on relationships between the information and an aspect of its use or context—were particularly problematic (Shreeves et al., 2005). Unlike intrinsic information quality dimensions in which the information can be measured in relation to a reference standard (such as a date encoding standard), measurement of relational quality dimensions are dependent on what context an item was meant for and its use within that context. In this environment, collection-level metadata could supply some of that context, given a better understanding of the relationships between collection and item level metadata.

In 2007 we received a new three year IMLS grant to continue the development of the registry and to explore how a formal description of collection/item metadata relationships could help registry users locate and use digital items. This latter activity, CIMR, (Collection/Item Metadata Relationships), consists of three overlapping phases. The first phase is developing a logic-based framework of collection/item metadata relationships that classifies metadata into categories with associated rules for propagating or constraining information between collection and item levels. Next we will conduct empirical studies to see if our conjectured taxonomy matches the understanding and behavior of metadata creators, metadata specification designers, and registry users. Finally we will design and implement pilot applications using the relationship rules to support searching, browsing, and navigation of the DCC Registry. We will also suggest OWL[26] bindings for the categories and inference rules. Although this framework will be applicable to collection-level descriptions generally, our initial focus is on the Dublin Core Collections Application Profile (DCMI, 2007).

The collection/item metadata relationships framework will allow metadata specification designers to more precisely indicate the relationships intended or assumed by their specifications. These applications of the framework are explicit classifications of metadata elements which will in turn provide guidance both to metadata creators assigning metadata and to systems designers mobilizing collection-level metadata in retrieval and browsing systems. In this way the framework supports:

- *Metadata specification developers defining metadata elements.* Metadata specification developers will be able to use applications of the framework to indicate the semantics of various metadata elements in their specifications.
- *Metadata creators describing objects.* Metadata librarians can use applications of the framework to confirm their understanding of the metadata elements they are assigning.
- *Systems designers developing and configuring retrieval systems.* Software architects can use applications of the framework to guide the design and implementation of automatic inferencing features in retrieval and browsing software.

In addition collection curators can use applications of the framework to improve metadata quality by discovering inconsistencies in metadata assignments between the collection and item levels, and to facilitate semantic interoperability with other databases and applications.

Many benefits of such a framework can be realized almost immediately. Later, when formal specifications and tools based on them are in place, the intended relationships (specified in a computer processable formats) can be integrated directly into management and use, as well as software. However realizing this level of value will require not only completing a plausible framework of relationships, but developing a public specification that is practical and reflects the

[26] http://www.w3.org/TR/owl-features/

common understandings of the metadata community. The current paper is only a first step in that direction. [27]

3. Three Kinds of Metadata Relationships

Currently we are focusing on defining categories for the simplest cases, where information recorded at the collection level can be usefully, if not always completely, converted to information at the item level. So far we have identified three categories, *attribute/value-propagation*, *value-propagation*, and *value-constraint*, which will serve to illustrate our approach.

Our characterizations are being developed in first order logic, extended as necessary by modal notions and other constructs. This is partly to ensure precision and clarity, and partly in anticipation of a final specification in RDF/OWL that will support automatic inferencing. However we work initially in first order logic rather than directly in OWL in order to take advantage of a compact familiar notation with well-understood semantics, and which can be easily extended as necessary to include modal, temporal, or other features. Since the use of first order logic with extensions will allow the expressiveness of our characterizations to be greater than that available in the appropriate level of OWL, a reductive strategy may be in order when we begin those translations.

3.1. Attribute/Value Propagation

Consider the DC Collections AP property *marcrel:OWN*, adapted from the MARC cataloging record standard. It is plausible that within many legal and institutional contexts whoever owns a collection owns each of the items in the collection, and so if a collection has a value for the *marcrel:OWN* attribute then each member of the collection will have the same value for *marcrel:OWN*. (For the purpose of our example it doesn't matter whether or not this is actually true of *marcrel:OWN*, only that some attributes are sometimes used by metadata creators with an understanding of this sort, while others, such as *dc:identifier*, are not). We refer to this meta-property of metadata elements as *attribute/value propagation* (or *a/v-propagation*). An informal definition might be:

Def a/v-p 1: an attribute **A** *a/v-propagates* =df
if a collection has some value *z* for **A**, then each item in the collection has *z* for **A**.

Some collection-level metadata elements a/v-propagate to collection members, and some don't — those that do present obvious opportunities to preserve context by bringing collection-level information to the item level.

A natural formalization of **Def a/v-p 1** in first order logic would be:

Def a/v-p 2: An attribute **A** *a/v-propagates* =df
$\forall x \forall y \forall z$ [(IsGatheredInto(x,y) & A(y,z)) $\supset$ A(x,z)]

Here we use *IsGatheredInto*, from the DCMI Collections AP to represent the item/collection relationship (DCMI, 2007). We assume that if something x IsGatheredInto something y then y is a collection and x is a member (of a collection). Or in the notation of first order logic: $\forall x \forall y$ [IsGatheredInto(x,y) $\supset$ (Member(x) & Collection(y))].

3.2. Interlude I: Propagation vs. Inheritance

Although attribute/value propagation from collection to members might be considered a kind of inheritance, in some very broad sense of inheritance, we think it is misleading to classify it as

[27] A briefer description of CIMR at an earlier stage of development is Renear et al. (2008a).

such. A little analysis shows that attribute/value propagation is in any event clearly not classical subsumptive inheritance as found in frame-based systems and semantic networks.

Consider a typical example of a taxonomic class hierarchy: Fido is an *instance of* the class DOG; DOG *is a subclass* of MAMMAL; and MAMMAL has the attribute/value pair *thermeoregulation=warmblooded*. DOG inherits *thermeoregulation=warmblooded* from MAMMAL in virtue of the fact that DOG is a subclass of (*a kind of*) MAMMAL; and that Fido inherits (although not in precisely the same sense) *thermeoregulation=warmblooded* from DOG because Fido is an instance of (*is a*) DOG. Note that there are two sorts of inheritance supporting relationships in our example: *subclass* and *instance*. The classical notion of inheritance has varying interpretations and ambiguities (Woods, 1975; Brachman, 1983), but in any case it is easy to see that neither of these two inheritance-supporting relationships, subclass and instance, matches the *IsGatheredInto* relationship between items and their collections: a member of a collection is neither a *subclass of* a collection nor an *instance of* that collection.

Our use of the term "propagation" in this sense is intended to follow Brachman (1991).

3.3. Value Propagation

Another collection/item metadata relationship is almost, but not quite, this simple. Consider the collection-level attribute *mycld:itemType*, intended to characterize the type of objects in a collection, with values from the DCMI Type Vocabulary (for the example we assume homogeneous collections, so this is an additional refinement on DCMI *cld:itemType*). Here we cannot conclude that if a collection has the value *dcterms:Image* for *mycld:itemType* then the items in that collection also have the value *dcterms:Image* for that same attribute. This is because an item that is an image is not itself a collection of images and therefore cannot have a value for *mycld:itemType*.

However, while the rule for propagating the information represented by *mycld:itemType* from collections to items is not simple propagation of attribute and value, it is nevertheless simple enough: if a collection has a value, say *dcterms:Image*, for *mycld:itemType*, then the items in the collection have the same value for a corresponding attribute, say, *dc:type*. The metadata elements *mycld:itemType* and *dc:type* have the same domain of values, but a different semantics. When two metadata attributes are related in this way we say the first *value-propagates* (or *v-propagates*) to the second. Informally:

Def v-p 1: an attribute **A** *v-propagates* to an attribute **B** =df
if a collection has the value *z* for **A**, then every item in the collection has the value *z* for **B**.

Notice that in this view, a/v-propagation is a special case of v-propagation: an attribute a/v-propagates precisely when it v-propagates to itself.

A formalization of **Def a/v-p 1** in the symbolism of first order logic would be:

Def v-p 2: An attribute **A** *v-propagates* to an attribute **B** =df
$\forall x \forall y \forall z$ [(IsGatheredInto(x,y) & $\mathbf{A}$(y,z)) $\supset$ $\mathbf{B}$(x,z)]

3.4. Value Constraints

Some collection/item metadata relationships are less direct than simple value propagation. In these cases, the value for the attribute on the item level is not the same, but does stand in some particular relation to the value for the collection-level attribute. For example, consider the collection-level attribute *mycld:dateItemsCreated* from the DC Collections AP, and the item-level attribute *mydc:created*. If a collection has a date range given as the value for *mycld:dateItemsCreated*, then we can infer about each item in that collection that a date given for the value of *mydc:created* will fall within that date range (for this example we assume neither of these attributes may be repeated, so these are again a refinement of the corresponding DCMI terms). We refer to these cases as *value constraints* (or *v-constraints*), since the collection-level metadata can be seen as constraining the values for a particular item-level attribute.

Informally:

Def v-c 1: an attribute **A** *v-constrains* an attribute **B** with respect to a constraint **C** =df if a collection has the value *z* for **A** and an item in the collection has the value *w* for **B**, then *w* is related to *z* by **C**.

The predicate variable **C** in the definition above represents the constraint between the values and will vary with the semantics of the related attributes. The constraint discussed in the example above is temporal containment, other sorts of constraints would be relevant to other sorts of metadata elements — for instance, spatial metadata might have spatial containment constraints. The modeling of this kind of metadata relationship may be useful for validation of item-level metadata in regard to the intent of the metadata creators.

A natural formalization for v-constraint would be:

Def v-c 2: an attribute **A** *v-constrains* an attribute **B** with respect to a constraint **C** =df

$$\forall x\forall y\forall z\forall w\ [(\text{IsGatheredInto}(x,y)\ \&\ \mathbf{A}(y,z)\ \&\ \mathbf{B}(x,w)) \supset \mathbf{C}(w,z)]$$

3.5. Interlude II: The Need for Modalization

Since the formalizations **Def a/v-p 2, Def v-p 2, and Def v-c 2** use truth-functional material conditionals ("P ⊃ Q") to express the conditional assertions seen in **Def a/v-p 1 Def v-p 1**, and **Def v-c 1** they fell prey to familiar difficulties sometimes referred to as the "paradoxes of material implication." The so-called paradoxes are the counterintuitive results that follow from the truth functional material conditional being defined as true whenever the antecedent is false (regardless of the truth value of the consequent), and whenever the consequent is true (regardless of the truth value of antecedent).

Consider the attribute, *acme:collIdentifier*, whose value is intended to be a collection identifier assigned by a particular identifier assignment agency, the ACME collection identifier agency. This attribute is obviously not a/v-propagating: one cannot conclude from the fact that a collection has a value for *acme:collIdentifier* that the items in the collection have that value (or even any value) for *acme:collIdentifier.* However before the assignment of any of these collection identifiers by the ACME agency there will be no collections with a value for *acme:collIdentifier*. Therefore, the conditional will be satisfied ("trivially") and *acme:collIdentifier* will be classified as a/v-propagating, which it is not.

To avoid this erroneous result, we can use a modal version of the conditional which, in the case of a/v-propagation, states that an attribute **A** a/v-propagates if and only if *it is impossible for*: a collection to have *v* for **A** and its items not have *v* for **A**.

Def a/v-p 2: An attribute **A** *a/v-propagates* =df

$$\Box\forall x\forall y\forall z\ [(\text{IsGatheredInto}(x,y)\ \&\ \text{A}(y,z)) \supset \text{A}(x,z)\ ]$$

Where the "□" is read "necessarily…".

However although this definition seems like a natural account of a/v propagation and does address the problem with attributes such as *acme:collIdentifier*, it still does not accurately identify all and only attributes that are (intuitively) a/v propagating. This is because modalized conditionals are themselves susceptible to a modal version of the paradoxes of material implication, sometimes called "the paradoxes of strict implication": if the antecedent of a modal conditional is *necessarily* false, then the conditional is true regardless of the consequent; and if the consequent is *necessarily* true, then the conditional is true, regardless of the antecedent. Our approach to this (also well-known) problem is to use preemptive modal restrictions to exclude the remaining counterexamples. A prose version of such a definition might be

Def a/v-p 4: An attribute **A** *a/v-propagates* =df
I. a) It is possible for a collection to have a value for **A**; &
b) It is possible for a collection member to have a value for **A**; &
c) It is possible that some value for **A** is had by one thing and
lacked by another; &
II. Necessarily, if some item is a member of a collection which has some
value for **A**, then that item has that value for **A**.

Or, in first order modal logic:

Def a/v-p 4: An attribute **A** *a/v-propagates* =df
I. a) $\Diamond\ \exists y \exists z\ [\mathrm{Collection}(y)\ \&\ \mathbf{A}(y,z)]$ &
b) $\Diamond\ \exists x \exists z\ [\mathrm{Member}(x)\ \&\ {\sim}\mathbf{A}(x,z)]$ &
c) $\Diamond\ \exists x \exists y \exists z\ [\mathbf{A}(x,z)\ \&\ {\sim}\mathbf{A}(y,z)]$ &
II. $\Box\ \forall x \forall y \forall z\ [(\mathrm{IsGatheredInto}(x,y)\ \&\ \mathbf{A}(y,z)\ ) \supset \mathbf{A}(x,z)\ ]$.

Where "◇" is read "it is possible that…" and is equivalent to "~□~", Similar modal definitions can be developed for v-propagates and v-constrains. For the rationale for these additional clauses see Renear et al. (2008b).

The problem of trivial satisfaction has been noted in the information retrieval literature, where van Rijsbergen (1986) and Lalmas (1998) argue that it is serious problem, and Sebastiani (1998) argues that it is not, claiming that the conditionals in question do not nest at the level where problems are created. Our analysis seems to support van Rijsbergen and Lalmas, at least for the applications being considered here. When conditionals are used in definitions, or in specification design and conceptual analysis, they do indeed nest at the problematic level, and in the problematic location (the definiens of a definition, or, more generally, in the antecedent of a larger conditional (when "=df" is read "if and only if").

Our particular solution to the problem, a combination of a modalized conditional and preemptive modal exclusion, suggests that any adequate representation of collection/item relationships will require modal notions. We note that our technique of modal exclusion is similar in some respects to the modal "metaproperty" strategy for ontology design (Guarino & Welty, 2004), where modal notions are also used to capture our intuitive understanding of fundamental concepts. We have discussed this problem in further detail elsewhere (Renear, et al., 2008b).

4. Future Research Directions

4.1. Extending the Framework

A complete framework for collection/item metadata relationships would cover not only the entailments from single assertions about collections to single assertions about items, but many other collection/item relationships.

Obviously one major division of collection/item metadata relationships is between those that support inferences from collection-level attributes to item-level attributes, and those that support inferences from item-level attributes to collection-level attributes. In this paper we have given examples of the former sort of relationship only.

Moreover, so far we have only considered cases where the assertion of a single metadata attribute at one level implied the assertion of a single metadata attribute at the other. But a complete framework for collection/item relationship categories must also accommodate the more general case, where assertions of one or more than one metadata attribute at one level imply assertions of one or more than one metadata attribute at the other level.

4.2. Intentionality

Throughout the discussion above we have carefully avoided directly raising questions such as "what is a collection?" and "what is it for something to be *gathered into* something else?". This is in part because we believe that answering those questions will necessarily involve the current analysis, and so consequently those questions are not genuinely prior, methodologically speaking, to our analysis of collection/item metadata relationships. In fact we see our analysis of collection/item metadata relationships as itself a substantive contribution to questions such as "what is a collection?". But in any case we cannot long avoid directly addressing the fundamental issue of the role of curatorial intent, which must be part of any analysis of the concept of a collection. When we do take up these issues directly it is quite likely that we will need to extend our logic further, to include intentional as well as alethic modal operators.

4.3. Reduction-Resistant Collection Level Properties

It would seem that some collection-level properties can be safely re-expressed as item-level metadata without loss of information. For instance, if a collection is described as being a collection of images we can (at least arguably) assume that nothing further is intended by that description than that each item in the collection is an image. In this case a/v-propagation and v-propagation carry all intended collection-level information to the item level and can straightforwardly support enhanced discovery and use.

However other sorts of collection-level information cannot be so easily reassigned to the item level without loss of meaning. In such cases the strategy of moving information from the collection level to the item level may still be valuable, but cannot, by itself, fully exploit the information provided at the collection level. Intriguingly these attributes often turn out to be carrying information that is tightly tied to the distinctive role the collection is intended to play in the support of research and scholarship. Obvious examples are metadata indicating that a collection was developed according to some particular method, designed for some particular purpose, representative in some respect of a domain, has certain summary statistical features, and so on. Such features cannot be converted to facts about individual items, and yet this is precisely the kind of information that makes a collection, as a *collection*, valuable to researchers — and if it is lost or inaccessible the collection cannot be useful in the way originally intended by its creators.

Understanding and exploiting metadata of this kind will be a particular challenge.

Acknowledgements

This research is supported by a 2007 IMLS NLG Research & Demonstration grant hosted by the GSLIS Center for Informatics Research in Science and Scholarship (CIRSS), Carole L. Palmer, Principal Investigator. Project documentation is available at http://imlsdcc.grainger.uiuc.edu/about.asp. We have benefited considerably from discussions with other DCC/CIMR project members and with participants in the IMLS DCC Metadata Roundtable, including: Timothy W. Cole, Thomas Dousa, Myung-Jan Han, Amy Jackson, Mark Newton , Carole L. Palmer, Sara L. Shreeves, Michael Twidale, Oksana Zavalina, and Wu Zhang.

References

Arms, William Y., Naomi Dushay, Dave Fulker, and Carl Lagoze. (2003). A case study in metadata harvesting: The NSDL. *Library Hi Tech, 21*(2), 228–237.

Brachman, Ronald J. (1983). What ISA is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer, 16* (10), 30-6.

Brachman, Ronald J., Deborah L. McGuinness, Peter F. Patel-Schneider, Lori A. Resnick, and Alex Borgida. (1991). Living with classic: When and how to use a KL-ONE-like language. In John Sowa, *Principles of semantic networks: Explorations in the representation of knowledge,* (pp. 401-456).

Brockman, William, Laura Neumann, Carole L. Palmer, Tonya.J. Tidline. (2001). *Scholarly work in the humanities and the evolving information environment*. Washington, DC: Digital Library Federation/Council on Library and Information Resources.

Christenson, Heather, and Roy Tennant. (2005). *Integrating information resources: Principles, technologies, and approaches*. California Digial Library. Retrieved from http://www.cdlib.org/.

Currall, James, Michael Moss, and Susan Stuart. (2004). What is a collection?. *Archivaria, 58,* 131-146.

Dempsey, Lorcan. (2005, October 9). From metasearch to distributed information environments. Message posted to http://orweblog.oclc.org/archives/000827.html.

DLF. (2005). The distributed library: OAI for digital library aggregation. *OAI Scholars Advisory Panel, 2005 June 20-21, Washington, DC.* Digital Library Federation.

DCMI. (2007). *Dublin Core Collections Application Profile*. Retrieved April 13, 2008, from http://dublincore.org/.

Dushay, Naomi, and Diane I. Hillmann. (2003). Analyzing metadata for effective use and re–use. *DC–2003: Proceedings of the International DCMI Metadata Conference and Workshop,* (pp. 161-170).

Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, and Sarah L. Shreeves. (2005). Using collection descriptions to enhance aggregation of harvested item-level metadata. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* (pp. 32-41). ACM Press.

Gasser, Less, Besiki Stvilia, Michael B. Twidale, and Linda C. Smith. (2001). *A new framework for information quality assessment*. Technical report ISRN UIUCLIS--2001/1+AMAS. Champaign, Ill.: University of Illinois at Urbana Champaign.

Guarino, Nicola, and Christopher A. Welty. (2004). An overview of OntoClean. In Steffen Staab and Rudi Studer, *Handbook on Ontologies*. Springer.

Heaney, Michael. (2000). *An analytic model of collections and their catalogues*. UK Office for Library and Information Science.

Hutt, Arven, and Jenn Riley. (2005). Semantics and syntax of Dublin Core usage in Open Archives Initiative data providers of cultural heritage materials. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, Colo, June 7–June 11,* (pp. 262–270). New York: ACM Press.

Lagoze, Carl, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom, and John Saylor. (2006). Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM Press.

Lalmas, Mounia. (1998). Logical models in information retrieval. *Information Processing and Management,* (*34)*1, 19-33.

Lee, Hur-Li. (2005). The Concept of Collection from the User's Perspective. *Library Quarterly, 75*(1), 67-85.

Lee, Hur-Li. (2000). What is a collection? *JASIS, 51*(12), 1106-1113.

Palmer, Carole L. (2004). Thematic research collections. In Susan Schreibman, Raymond G. Siemens, and John Unsworth (Eds.), *Companion to digital humanities* (pp. 348-365). Oxford: Blackwell Publishing.

Palmer, Carole L., and Ellen M. Knutson. (2004). Metadata practices and implications for federated collections. *Proceedings of the 67th ASIS&T Annual Meeting.*

Palmer, Carole L., Ellen M. Knutson, Michael Twidale, and Oksana Zavalina. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting, Austin, Texas, 2006.*

Renear, Allen. H., Richard Urban, Karen Wickett, C. L. Palmer., and David Dubin. (2008a). Substaining collection value: Managing collection/item metadata relationships. *Proceedings of the Digital Humanities conference, Oulu, Finland, 2008.*

Renear, Allen. H., Richard Urban, Karen Wickett, and David Dubin. (2008b). The return of the trivial: Formalizing collection/item metadata relationships. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries 2008*. New York: ACM Press.

Sebastiani, Fabrizio. (1998). On the Role of Logic in Information Retrieval. *Information Processing and Management 34*(1), 1-18.

Shreeves, Sarah L. , Ellen M. Knutson, Besiki Stilva, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole. (2005). Is 'Quality' Metadata, 'Shareable' Metadata? The Implications of local metadata practices for federated collections. *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, Minneapolis, 2005, (*pp. 223-237). Chicago, IL: Association of College and Research Libraries.

Stvilia, Besiki, Less Gasser, Michael B. Twidale, Sarah L. Shreeves, and Thomas W. Cole. (2004). Metadata quality for federated collections. *Proceedings of ICIQ 04-9th International Conference on Information Quality, Cambridge, MA*.

Van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. *The Computer Journal 29*(6), 481-485.

Warner, Simon., Jeroen Bekaert, Carl Lagoze, Xiaoming Lin, Sandy Payette, and Herbert Van de Sompel. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries.* doi: 10.1007/s00799-007-0016-7

Wendler, Robin. (2004). The eye of the beholder: Challenges of image description and access at Harvard. In Diane I. Hillmann, and Elaine L. Westbrooks (eds.), *Metadata in Practice,* (pp. 51-6). Chicago: American Library Association.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In Daniel G. Bobrow, and Allan Collins (eds.), *Representation and Understanding: Studies in cognitive science*. New York: Academic Press.

Full Papers

Session 4:

Metadata Quality

Answering the Call for more Accountability: Applying Data Profiling to Museum Metadata

Seth van Hooland
ULB, Belgium
svhoolan@ulb.ac.be

Yves Bontemps
IBM, Belgium
yves.bontemps@be.ibm.com

Seth Kaufman
OpenCollection, USA
seth@opencollection.org

Abstract

Although the issue of metadata quality is recognized as an important topic within the metadata research community, the cultural heritage sector has been slow to develop methodologies, guidelines and tools for addressing this topic in practice. This paper concentrates on metadata quality specifically within the museum sector and describes the potential of data-profiling techniques for metadata quality evaluation. A case study illustrates the application of a general-purpose data-profiling tool on a large collection of metadata records from an ethnographic collection. After an analysis of the results of the case-study the paper reviews further steps in our research and presents the implementation of a metadata quality tool within an open-source collection management software.

Keywords: metadata quality; data-profiling; collection management software

1. Introduction

Collection registration technologies for cultural heritage resources have greatly improved during the last three decades, gradually transforming card catalogs to web-based applications. Successive technologies have impacted the content of both newly created metadata and existing metadata migrated from older platforms. A good example of the influence of a specific technology on content is the character field length limitations of punch cards fed into mainframes in the 1970's, the effects of which are still felt today in some legacy data sets. Technological evolutions have also been accompanied by (and partially engendered) a shift in the profile of professionals working with these tools to document collections. There is, for example, a clear tendency within cultural institutions to give the repetitive work of metadata creation to administrative and technical staff, apprentices or student workers, whereas collection description used to be performed by specifically trained staff members. In multi-lingual countries such as Belgium one also has to consider the complexity of collections being described sometimes in one language, sometimes in another, depending on the mother tongue of the staff. Under these circumstances vast repositories of metadata records have been created and migrated from one platform to another, with little or no information regarding their consistency, completeness and accuracy.

As long as the metadata remained within the safe boundaries of the museum this was not such a problem. Users submitted their question to a member of the museum staff that could query the database for them. As such, the database (and the metadata records it contained) was more or less treated as an internal tool. But then came the web. Initially, most museum web-presences were limited to basic institutional information. Only a very limited number of museums published their metadata in the same way as libraries, which offered their users an OPAC. But the growing tendency to aggregate thematically or geographically related metadata from libraries, archives and museums with the use of OAI-PMH has raised the pressure on museums to publish or distribute all of their available metadata. The disappointing quality of search results and the minimal descriptions attached to retrieved objects within such projects has led to a discussion on issues surrounding the consistency, accuracy and completeness of metadata.

This discussion is badly needed as collection holders increasingly try to re-use metadata and gain more value from them within digitization projects. Metadata practitioners assisting digitization projects that aggregate metadata of different partners must acknowledge that the quality of existing metadata is hardly questioned. After all, which collection holder wants to stand up in the middle of his or her peers and warn them about the poor quality of his or her metadata?, This misplaced trust causes delays and failures when metadata do not live up to expectations. But more importantly, the community must acknowledge the lack of established standards, methodologies or tools for metadata quality evaluation. Or to put it in the often-cited words of Diane Hillmann: "There are no metadata police".

In the absence of such standards or tools metadata practitioners usually believe that documenting the quality of their metadata is too costly a project to be undertaken. This paper shows that useful metadata indicators can be produced at a very low cost from existing metadata using general-purpose data-profiling tools. In order to facilitate the measurement and improvement of metadata we propose to integrate such tools with collection management applications, making quality measurement a continuous and seamless task. This will remove the barriers that currently prevent practitioners from actually acting on issues of metadata quality.

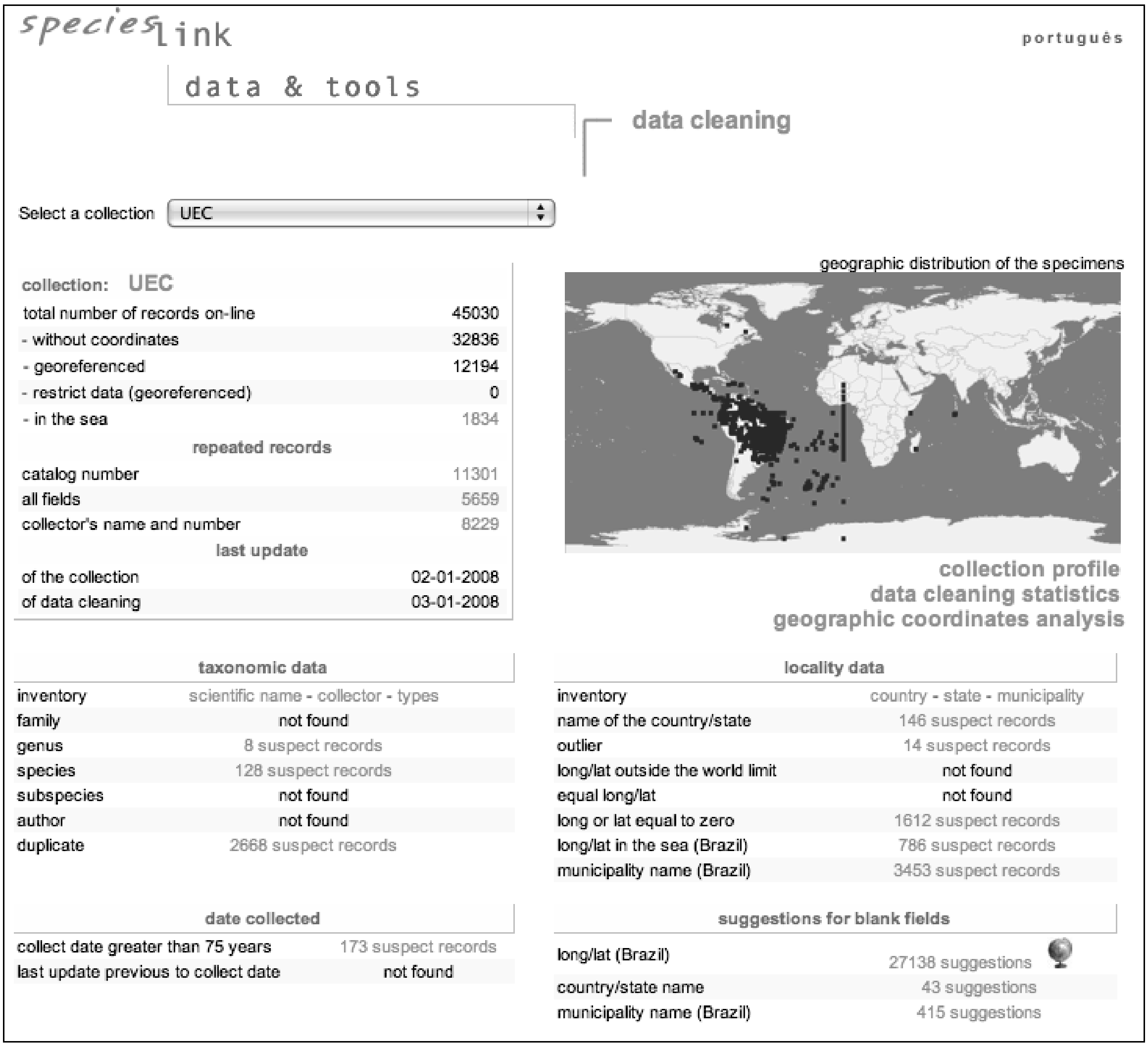
2. Overview of the Research

2.1. Global Data Quality Research

Metadata quality is, obviously, not only an issue for the cultural heritage sector. A large body of research, development and tools has been developed throughout the 1990's within the computer science field, the corporate world and public administrations to examine the notion of data or information quality. A multitude of other denominators and sub activities, such as data cleaning, -profiling and –standardization exist. An overview of the data quality field can be found in "Data quality: concepts, methodologies and techniques" by Batini and Scannapieco (2006) and "Data Quality : the Field Guide (2001) by Thomas Redman.

Within this large domain it is the specific topic of data profiling that is of special interest to us. Data profiling is the first and the essential step towards data quality in the sense that it consists of gathering factual information on the data quality that can be used, firstly, to decide which actions to take in order to enhance quality and, secondly, to inform users about the quality of the data they are consulting. An automated implementation of a data profiling procedure could reduce uncertainty and misconceptions regarding the quality of our collection registration databases. Collection managers and the public alike sorely need concise reports consisting of up-to-date statistical information on the quality of the totality of the records.

The application and utility of such a tool can be demonstrated by taking a look at another domain. An interesting application that might inspire methodologies and tools for the cultural heritage sector is offered by the research community around biodiversity data. The aggregation of huge sets of scientific data concerning climate, flora and fauna resulted in the same problems mentioned above. The Reference Center for Environmental Information of Brazil therefore has developed a data cleaning tool which aims to help curators identify possible errors. The system presents "suspect" records, recommending that they be checked by the author or curator.



Select a collection UEC

collection: UEC	
total number of records on-line	45030
- without coordinates	32836
- georeferenced	12194
- restrict data (georeferenced)	0
- in the sea	1834
repeated records	
catalog number	11301
all fields	5659
collector's name and number	8229
last update	
of the collection	02-01-2008
of data cleaning	03-01-2008

taxonomic data	
inventory	scientific name - collector - types
family	not found
genus	8 suspect records
species	128 suspect records
subspecies	not found
author	not found
duplicate	2668 suspect records

locality data	
inventory	country - state - municipality
name of the country/state	146 suspect records
outlier	14 suspect records
long/lat outside the world limit	not found
equal long/lat	not found
long or lat equal to zero	1612 suspect records
long/lat in the sea (Brazil)	786 suspect records
municipality name (Brazil)	3453 suspect records

date collected	
collect date greater than 75 years	173 suspect records
last update previous to collect date	not found

suggestions for blank fields	
long/lat (Brazil)	27138 suggestions
country/state name	43 suggestions
municipality name (Brazil)	415 suggestions

FIG. 1: Screenshot of a data cleaning tool from the biodiversity domain (http://splink.cria.org.br/dc/)

Figure 1 represents information that is generated on the fly on the actual data by pointing out how many records are online, how many of them are geo-referenced, how many duplicated records have been detected, when the last update of the collection took place, etc. Each time suspect records are mentioned a direct link is provided to verify manually in detail the record and its metadata. Among the options offered on this page we especially would like to point out the possibility to visualize the data cleaning statistics as graphs representing the evolution through time of the number of suspect authors, duplicated records and catalog numbers (see FIG. 2).

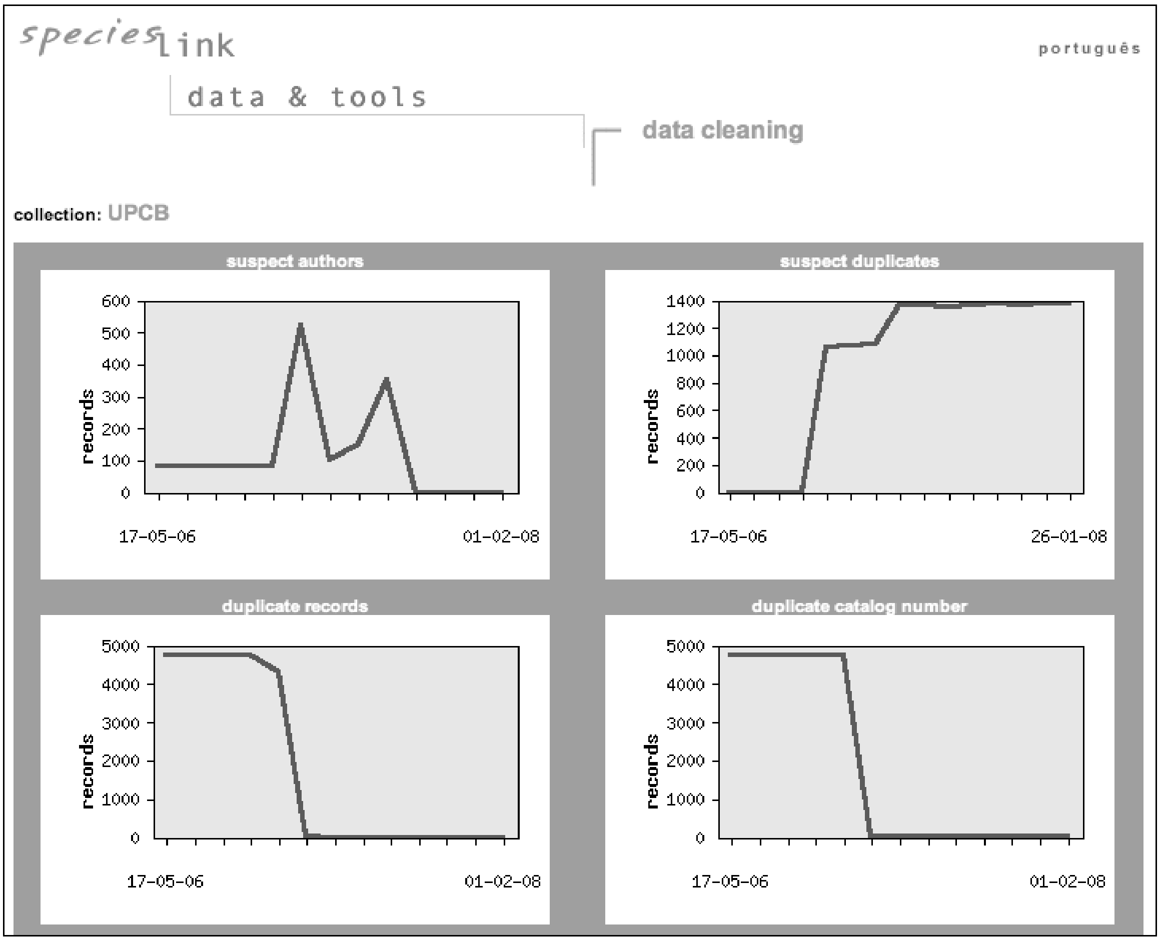


FIG. 2: Graphs representing the evolution of data quality within the biodiversity domain (http://splink.cria.org.br/dc/)

This tool offers the opportunity for a potential user of the collection to grasp within ten or fifteen minutes the quality of the data he or she is interested in.

2.2. Specificity of the Cultural Heritage Sector

Now that we have given an example from another application domain we should try to define the specific problems and characteristics related to the cultural heritage sector in order to see how tools from other domains could be applied to museum metadata.

Firstly, in contrast with information systems from other domains, such as the financial or the administrative sectors, the direct economic value of the metadata from the cultural heritage sector are comparatively limited. Metadata could play a crucial role in the re-use and marketing of digital cultural heritage, but European reports and projects investigating business models based on the commercialization of digital cultural heritage from the public domain do not point to viable options. Put simply, one cannot expect a traditional return on investment of digitization projects in the sense that the market validation of digital cultural heritage is not likely to make up for the investments made for the digitization. But this does not mean the sector cannot learn something from more economically viable domains, where data-profiling tools offer a means to introduce more accountability through statistical monitoring. The public financing of long-term metadata creation projects is unfortunately sometimes regarded as throwing money into a black hole. Data profiling could help to quantify the efficiency and effectiveness of metadata creation throughout the project life-cycle.

Secondly, museums and other heritage institutions often find it hard to define the exact needs of their users, especially when the collections consist of art. Compared to other application domains, user needs regarding cultural heritage are mostly defined in very general and vague terms. This makes metadata evaluation difficult since quality is at its most abstract level defined as the "fitness for purpose". But how can this be judged without sufficient knowledge of user expectations? Log files of user queries are haphazardly used for research purposes (Cunningham and Mahoui, 2000), but the logs have little real impact on collection description. Recent experiments with user-generated metadata such as user comments and folksonomies offer an interesting step in this direction (van Hooland, 2005). In a broad sense logs of user queries, comments and tags could also be considered as metadata linked to the collection, to which data profiling can be applied in order to more easily detect patterns and recurrences.

Lastly, we must to point out the empirical and non-structured character of cultural heritage documents. It is the core-business of heritage holders to manage and facilitate access to historical collections, for which it can be very time-consuming and sometimes impossible to document the origin and intent of the collection. Sometimes old documentation can exist, but the updating of legacy metadata is a necessity. This illustrates the problem of the ever-extendibility of metadata, in the sense that metadata themselves have to be documented as the reality and its definition evolve throughout time. But administrative or legislative institutions, which are obliged to retain their historical data, are also confronted with shifting definitions, domains and attributes of concepts such as, for example, unemployment, nationality or retirement (Boydens, 2001). The unstructured character of cultural heritage information is also blamed for the difficulty of inserting documentation into rigorously defined database fields. The extensive work in recent years on metadata models has attempted to structure as much as possible the documentation process by providing clear-cut definitions of metadata fields and sets of possible values (e.g with controlled vocabularies). But still, the descriptions that contain key information for users are contained in free-text fields. It is precisely the automated analyses of unstructured text which poses problems when assessing metadata quality.

2.3. Current Research within the Cultural Heritage Sector

The first discussions on metadata quality within the cultural heritage sector dealt with bibliographic control in the library world. However, the growing variety of types of resources, their metadata formats and user communities called for an enlarged scope. Bruce and Hillmann (2004) provide the first major theoretical foundation regarding metadata quality with their “systematic, domain- and method-independent discussion of quality indicators”.

Defining quality measurements and metrics is essential, but they also have to be put into practice. The manual analysis of a limited sample of the complete set of metadata records has been a way to gather interesting indications (Shreeves et all, 2005). However, this manual approach has two obvious disadvantages: 1) it is too time consuming (and thus too expensive) and 2) it only offers a “photograph” of a sample of the metadata records at one specific moment in time. Therefore, we will focus only on practical semi-automated approaches that can repeatedly analyze the totality of a given metadata set.

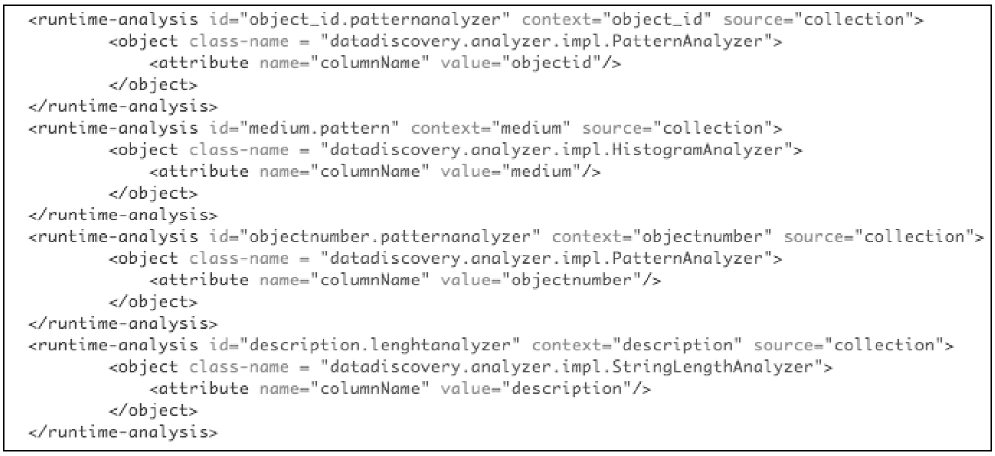
Tennant (2004) proposes a minimal, pragmatic set of analysis functions to be applied on metadata and specifies queries to be computed such as the total number of occurrences of a certain value or patterns across records (e.g. all records with “x” in the “y” field do not have a “z” field). The application of such scripts or queries on large numbers of metadata records produces results which are difficult to grasp without the aid of visualization software. Dushay and Hillmann present a tool that can translate the results of queries upon a large collection of records into a human-readable form that allows the detecting of patterns and the extent of the problems (Dushay and Hillmann, 2003). Several researchers have also worked on metadata transformation and enrichment, especially in the context of aggregated content projects. Foulonneau and Cole (2005) report, for example, on how harvested records can be transformed to be of higher use in the context of an OAI service provider.

Automated quality assessment normally concentrates on what in French is referred to as *critique externe* in the context of the evaluation of historical sources: it focuses on the formal characteristics of metadata, and not on its actual content. The *critique interne* is left to human evaluation, since it is impossible to develop automated tools to grasp evaluation criteria such as accuracy and conformance to expectations. Ochoa and Duval (2007) however propose to translate these and the other criteria from the Bruce and Hillmann framework into equations that can be automatically applied. Still, this approach only applies to metadata of textual resources and not to other types of unstructured data such as images.

One of the most promising ideas has been formulated by Hillmann and Phipps (2007) who advocate the machine readability of application profiles. The real power of these “templates for expectation” can only be unleashed if their statements can be matched with the actual syntax and content of the metadata in an automated manner. But the automated validation of XML and RDF that wants to go further then just checking the “well-formedness” is still problematic, even though progress is being made (Brickley 2005).

3. Applying Data Profiling Techniques to Museum Metadata

Most of the research mentioned above used custom-written queries to be applied to the metadata records. This paper explicitly proposes to use a data profiler. Olson (2002) defines data profiling as “the use of analytical techniques to discover the true structure, content, and quality of a collection of data”. We are interested to see which results can be obtained by using an open-source general-purpose data profiling tool, available at http://sourceforge.net/projects/dataprofiler/ that works in three steps. First, the analysis to perform on the dataset has to be set up by creating an XML profile specification file (see figure 3) in which is specified which analysis runs on which column of the dataset. Five analyses are at our disposal, which we will present with the help of examples from our test collection. In a second step, the profiler itself is launched, which will read the XML file and store the result of the profiling into a local repository and the information about the profiling execution into a catalog file. The catalog file is used to record what profile specification (.xml file) was used as a basis for profiling and to retrieve the results from the local repository. Third, the visualizer is run to view the profile execution results. These can then be exported for further analysis in other tools.



```
<runtime-analysis id="object_id.patternanalyzer" context="object_id" source="collection">
        <object class-name = "datadiscovery.analyzer.impl.PatternAnalyzer">
            <attribute name="columnName" value="objectid"/>
        </object>
</runtime-analysis>
<runtime-analysis id="medium.pattern" context="medium" source="collection">
        <object class-name = "datadiscovery.analyzer.impl.HistogramAnalyzer">
            <attribute name="columnName" value="medium"/>
        </object>
</runtime-analysis>
<runtime-analysis id="objectnumber.patternanalyzer" context="objectnumber" source="collection">
        <object class-name = "datadiscovery.analyzer.impl.PatternAnalyzer">
            <attribute name="columnName" value="objectnumber"/>
        </object>
</runtime-analysis>
<runtime-analysis id="description.lenghtanalyzer" context="description" source="collection">
        <object class-name = "datadiscovery.analyzer.impl.StringLengthAnalyzer">
            <attribute name="columnName" value="description"/>
        </object>
</runtime-analysis>
```

FIG. 3: Illustration of the XML profile specification file

We have tested the profiler on a comma-delimited export file from the ethnographic department of the Royal Museum for Central Africa consisting of 69,719 records, each record consisting of 13 fields (object id, object number, object count, date of collection, date of entry, date of production, title, medium, dimensions, thesaurus terms, description, old region, actual region). The majority of the metadata are in French, with Dutch being used in a few cases.

The end result of the profiling process is the creation of a report which specifies for each metadata field a rigorous definition, the domain the values can belong to and the referential

integrity rules with other metadata fields. Results of the different analyses allow the analyst to discover violations against the definition, domain and referential integrity rules of each metadata field. We will now illustrate the different analyses with examples from our test collection.

3.1. NullCount Analysis

The NullCount analysis calculates the number of records where the specified column holds no value. Table 1 illustrates the high number of records that have no value for certain fields. Several fields, such as "description", "dimensions", "date_of_production", "date_of_collecting" and "creditline" have no value 90% of the time, which is cause for concern. Users expect values in fields, especially fields as basic as 'description.

TABLE 1: Percentage of empty fields

Fieldname	**Percentage of empty fields**
objectid	0%
objectnumber	0%
objectcount	0%
date_of_collecting	87,5%
date_of_entry	55,6%
date_of_production	92%
title	8%
medium	66.3%
dimensions	90.7%
creditline	89.5%
description	92.7%
region_old	44%
region_new	44%

3.2. Pattern Analysis

The Pattern analysis calculates the different formats used to represent values. The values can be alphabetical characters (represented by the profiler with A), numerical characters (represented by the profiler with 9) or other special signs such as a punctuation sign or a slash. This analysis is particularly useful to examine the values that correspond to a certain fixed syntax, such as accession numbers and dates. The accession number in the case of our data set has to correspond to the following fixed syntax: [collection code].[inscription year].[lot number].[number of the item within a lot]-[number that indicates that the item is a part of series]. When running the pattern analyzer, we can see that 92% of the values match the required syntax.

The different date fields also offer an excellent opportunity to apply the pattern analysis. There is a total number of 52 different ways to encode the date_of_collecting. This is due to the fact that other information is also saved within the field in some cases. Obviously, this practice should be avoided. Table 2 represents the 10 most frequent patterns used to represent the date when an item was acquired and clearly demonstrates the need to standardize the input of dates.

TABLE 2: the 10 most recurrent patterns for the date_of_collecting field.

Pattern	Number of occurrences	Example
(empty)	65011	
9999-9999	1564	1891-1912
9999	1105	1909
99-99/9999	574	09-10/1992
99/9999	347	01/1994
99-9999	346	08-1950
99/99/9999	312	04/08/1963
AAA 9999	90	Mai 1938
AAAAAAA-AAAA 9999	84	Janvier-mars 1999
99-99 9999	61	01-02 1993

The same conclusion can be drawn from the results of the pattern analysis when applied on the dimension field (see table 3). Measures are not standardized (both mm and cm are used) and apparently no rules were laid down regarding the syntax. As in the case of the problem with dates, this incoherence makes the searching difficult, not to say completely impossible. The output of this type of analysis can be used to develop scripts for normalization and to build up value vocabularies.

TABLE 3: examples of different patterns to describe dimensions.

Pattern	Number of occurrences	Example
99 A 99 AA	1190	13 x 18 cm
999 AA	388	920 mm
999 A 999	382	573 x100
99 AA A 99AA	196	37 mm x 16 mm
99 AA A 99 AA A 99 AA	107	52 cm x 25 cm x 25 cm
99	14	72

3.3. Histogram Analysis

The histogram analysis produces a histogram of the different values that exist for a specific metadata field. We can apply this analysis to quite a range of fields. Table 4 represents for example the titles that appear more than a thousand times throughout the collection. These data can serve as an excellent guide for discussions regarding the precision of the terms used in fields.

TABLE 4: Most frequent titles.

Title	Number of occurrences
(empty)	5623
statuette	2043
panier	1800
bracelet	1792
collier	1376
masque	1324
groupe	1250
couteau	1073
sifflet en bois	1012

“By accident” strange values may be discovered by this analysis. For example, when applied to the field “object_count” the histogram analysis shows us that 39 fields have the value “0”, which is a violation of domain range integrity since an object must at least consist of one item.

3.4. Case Analysis

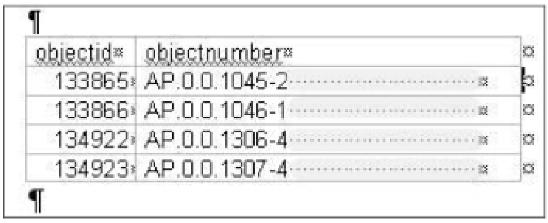
The case analysis gives an overview of the use of capitalized and non-capitalized alphabetic characters. The application of this analysis is rather limited but still enables one to check the level of consistency of the metadata input.

TABLE 5: Use of upper- and lowercase characters.

Case type	**Number of occurrences**	**Frequency (on the total number of non-empty fields)**
Mixed case	21186	54.7%
All uppercase	14889	38.4%
All lowercase	2645	6.8%

3.5. Length Analysis

The length analysis calculates the number of characters used in a field. Again, this is a very basic query that is performed on the metadata but its application can lead to interesting and unexpected results. When applied to the field “objectnumber”, the profiler informs us that 69,718 values consist of 42 characters and one value consists of 55 characters, although we see that the format of this field varies and never takes up 42 characters. The most frequent pattern “AA.9999.99.99” only consists of 13 characters, so where do these values come from? Figure 9 shows the reason behind these values. A copy/paste of the data within a text editor such as Word reveals the formatting of the characters and explicitly shows the whitespaces that are included within each value. The same phenomenon appears for the field “date_of_production”. Although the waste of storage space within the database is perhaps no longer a critical issue, the discrepancy between how the values are perceived and their true composition can poses problems for the long-term preservation of the metadata.



objectid	objectnumber
133865	AP.0.0.1045-2
133866	AP.0.0.1046-1
134922	AP.0.0.1306-4
134923	AP.0.0.1307-4

FIG 4: Presence of whitespaces within values.

4. Research and Development Agenda: Internalizing Metadata Quality within the Creation Workflow

The different analyses illustrated above clearly prove that simple and inexpensive data profiling techniques can bring many problems or particularities within large sets of metadata to the surface quite easily. But applying external tools on a periodic basis remains too much an ad-hoc solution to serve as an effective management tool for metadata quality improvement activities. And just as with manual sampling methods it only produces a “photograph” of the state of the metadata records at a specific moment in time. Ochoa and Duval (2007) point out a soft spot when they refer to metadata quality analysis as a “research activity with no practical implications in the functionality or performance of the digital repository.”

The only way to effectively have a day-to-day impact on metadata quality is to seamlessly implement a data profiling procedure within the metadata creation workflow. In the context of museum metadata the collection management system should thus incorporate functionality that enables collection managers to automatically draw data profiling reports with statistics and graphs that enable the continuous monitoring of the evolution of metadata quality.

No existing software offers such functionality. Therefore, we have established a collaboration with the development team of the open-source collection management software OpenCollection to develop and implement a metadata quality tool within that software package. OpenCollection is a general-purpose collection management system intended for use with a wide variety of materials. Current users include representatives from many fields, including fine art, anthropology, film, oral history, local history, architecture, material culture, biodiversity conservation, libraries, corporate archives and digital asset management. The most important features concerning metadata management are :

1. Completely web-based user interface, meaning that metadata input can be very easily distributed among a large group of indexers/catalogers or external experts.
2. Configurable, type-specific user defined key/value attribute system. In addition to the standard set of OpenCollection fields representing concepts applicable to anything that can be cataloged — things like "accession number" — sets of attributes functioning as repeatable custom fields,) may be defined. These sets can map to established metadata standards such as Dublin Core, Darwin Core, VRA Core 3.0, CDWA Lite, et. al. Attribute sets may be type-specific: they can be defined such that they are only available for specific types of cataloged items (ex. photographs, video tapes, films). They may also be repeating, and it is possible to impose an intrinsic data type (text, integer or floating point number, date) as well as bounds and pattern-based input validation rules.
3. Automatic extraction of metadata from uploaded media files.
4. Extensive support for authority lists and controlled vocabularies. A tool is included to import Getty Art and Architecture Thesaurus (AAT) data files.

We are currently evaluating several strategies for integration of the metadata quality tools described in this paper with OpenCollection. These range from straightforward inclusion of metrics generated by our tool in OpenCollection’s reporting system to more interactive approaches built into the metadata creation workflow itself. Examples of the latter include:

1. Dynamic evaluation during input of attributes, with display of quality/suitability metrics and, when possible, suggestions for improvement.
2. Visible per-record and per-field indicators of measured quality. The indicators are color coded and can provide detailed quality metrics on-demand.
3. Expansion of the OpenCollection search engine to support searches on quality metrics. Metric search criteria may be freely mixed with traditional content-based search terms, enabling users to efficiently locate groups of related problematic data.

The seamlessly integrated metadata quality module would be packaged with analyses available out-of-the-box. This would allow metadata practitioners to have a clear view on the state of their metadata. Hopefully, getting this first “general” summary for free will catch their attention to the metadata quality issue and drive them to improve quality.

5. Conclusions

This article has given a concise overview of the metadata quality issue and its specific nature within the cultural heritage sector. Secondly, a general-purpose data-profiling tool has been applied to a large test-collection of museum metadata which resulted in the identification of various problems and particularities in the metadata. Taking these results a step further we are finally promoting a pro-active way of dealing with metadata quality by endeavoring to directly incorporate a methodology and tool in an open-source collection management system. This innovative approach will introduce more accountability into the metadata creation process as a whole, which is at the moment all too often considered as a form of black art.

Acknowledgements

The authors would like to thank Hein Vanhee of the Royal Museum for Central Africa for supplying the test-data and the students of the course "Qualité de l'information et des documents numériques" of Professor Isabelle Boydens and Yves Bontemps of the Information and Communication Science Department of the Université Libre de Bruxelles, for performing analyses on the test-data.

References

Batini, Carlo, and Monica Scannapieco. (2006). *Data quality: Concepts, methodologies and techniques*. New York: Springer.

Boydens, Isabelle. (2001). *Informatique, normes et temps*. Bruxelles: Bruylandt.

Brickley, Dan. (2005). *CheckRDFSyntax and Schemarama Revisited.* Retrieved May 20, 2008, from http://danbri.org/words/2005/07/30/114.

Bruce, Thomas, and Diane Hillmann. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In Diane I. Hillmann & Elaine L. Westbrooks (Eds.), *Metadata in practice*, (pp. 238-256). Chicago: American Library Association.

Cunningham, Sally Jo, and Malika Mahoui. (2000). A comparative transaction log analysis of two computing collections. *4th European Conference on Digital Libraries*, (pp. 418-423).

Dushay, Naomi, and Diane Hillmann. (2003). Analyzing metadata for effective use and re-use. *DC-2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications, September 28 - October 3, 2003.* Retrieved February 14, 2008, from http://www.siderean.com/dc2003/501_Paper24.pdf.

Hillmann, Diane, and Jon Phipps. (2007). Application profiles: Exposing and enforcing metadata quality. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, *Singapore*. Retrieved May 20, 2008, from http://www.dcmipubs.org/ojs/index.php/pubs/article/viewFile/41/20.

Foulonneau, Muriel, and Timothy Cole. (2005). Strategies for reprocessing aggregated metadata. *Lecture notes in computer science 3652,* (pp. 290-301). Berlin, Heidelberg: Springer, ECDL. Retrieved February 14, 2008, from http://cicharvest.grainger.uiuc.edu/documents/metadatareprocessing.pdf.

Ochoa, Xavier, and Erik Duval. (2007). Towards automatic evaluation of metadata quality in digital repositories. *Ariadne.* Retrieved February 14, 2008, from http://ariadne.cti.espol.edu.ec/M4M/files/TowardsAutomaticQuality.pdf.

Olson, Jack. (2002). *Data quality: The accuracy dimension*. San Francisco: Morgan Kaufman.

Redman, Thomas. (2001). *Data quality:The field guide*. New Jersey, Boston: Digital Press.

Shreeves, Sarah, Ellen Knutson, Besiki Stvilia, Carole Palmer, Michael Twidale, and Timothy Cole. (2005). Is « quality » metadata « shareable » metadata ? The implications of local metadata practices for federated collections. *ACRL Twelfth National Conference*. Minneapolis: ALA.

Tennant, Roy. (2004). *Specifications for metadata processing tools.* 2007, 1(2), California Digital Library. Retrieved February 14, 2008, from http://www.cdlib.org/inside/projects/harvesting/metadata_tools.htm.

Van Hooland, Seth. (2005). Spectator becomes annotator: Possibilities offered by user-generated metadata for image databases. *Proceedings CILIP Cataloguing & Indexing Group Annual Conference, University of East Anglia, UK, 13-15 September 2006.* Retrieved February 14, 2008, from http://homepages.ulb.ac.be/~svhoolan/Usergeneratedmetadata.pdf.

A Conceptual Framework for Metadata Quality Assessment

Thomas Margaritopoulos
University of Macedonia, Greece
margatom@uom.gr

Merkourios Margaritopoulos
University of Macedonia, Greece
mermar@uom.gr

Ioannis Mavridis
University of Macedonia, Greece
mavridis@uom.gr

Athanasios Manitsaris
University of Macedonia, Greece
manits@uom.gr

Abstract

Metadata quality of digital resources in a repository is an issue directly associated with the repository's efficiency and value. In this paper, the subject of metadata quality is approached by introducing a new conceptual framework that defines it in terms of its fundamental components. Additionally, a method for assessing these components by exploiting structural and semantic relations among the resources is presented. These relations can be used to generate implied logic rules, which include, impose or prohibit certain values in the fields of a metadata record. The use of such rules can serve as a tool for conducting quality control in the records, in order to diagnose deficiencies and errors.

Keywords: digital repositories; metadata quality; related resources; logic rules

1. Introduction

The quality of metadata describing digital resources stored in a repository can be considered as a necessary condition for reliable and efficient operation of the repository. Metadata is considered to be the key to successfully discovering the appropriate resources. Therefore, metadata must be created and maintained according to well-defined procedures. This requirement is more important considering the vast number of available digital resources, which keeps on growing with rapid rates. Even though the requirement for quality metadata has been generally recognized, there isn't any commonly accepted approach on the definition of metadata quality, and, as a consequence, on the ways this quality can be assessed, measured and increased.

Studies conducted on the subject, represent research efforts to compute statistical indices (Najjar, Ternier & Duval, 2003; Friesen, 2004; Bui & Park, 2006), define frameworks (Moen, Stewart & McClure, 1997; Gasser & Stvilia, 2001; Bruce & Hillman, 2004), identify quality characteristics and detect quality problems (Dushay & Hillman, 2003), either directly or indirectly (by locating indicators of quality). The diversity and complexity of the proposed parameters or characteristics of metadata quality brings out the obvious need to return back to the basics and talk about the roots of the issue of quality and its fundamental components. A conceptual framework to define metadata quality by using analogies from common knowledge and experience is among the goals of this paper.

Moreover, an important conclusion drawn from studying relevant research efforts is that the majority of them assess quality of a metadata record or a metadata repository based on the syntactical level of the content and the metadata standard, but not on the semantical level. A potential source of semantical level information could be any possible interdependencies connecting the resources. Digital resources stored in a repository are not completely independent from each other; they are connected with structural or semantical relations. Especially, in digital resources constituting assemblies (like educational resources registered in a repository as collections, e.g. SCORM), or aggregations (e.g. a web page containing an image and an animation) these relations among the resources create a net of interdependencies, which affect their metadata records, accordingly. These interdependencies are expressed as logic rules the validity of which influences metadata quality and will be dealt with in this paper.

The rest of the paper is structured as follows: In Section 2, a literature review on the related work on the general subject of metadata quality, along with the subject of logic rules connecting metadata of related resources is conducted. In Section 3, a conceptual framework of metadata quality originating from an intuitive and empirical metaphor is proposed. Based on the framework introduced in Section 3, Section 4 presents a method of metadata quality assessment that uses logic rules connecting related resources. Section 5 provides application examples on the way such rules can be used to assess metadata quality. Finally, Section 6 draws conclusions and points out issues for future work.

2. Related Work

The related work presented in this section concerns different fields of study, which are combined for the purpose of the proposed approach; the field of metadata quality and the field of logic rules involving metadata of related resources.

In the past, several research efforts related with metadata quality have been conducted. These efforts approach the subject from diverse perspectives, trying to cover most of its different aspects. Najjar, Ternier & Duval (2003), Friesen (2004) and Bui & ran Park (2006) conduct a statistical analysis on a sample of metadata records from various repositories and evaluate the usage of the standard. They designate the most frequently used fields and values attributed to these fields. While not directly associated with quality, the statistical indices produced provide an insight of the efficiency of the repositories examined. In this regard, (Greenberg et al., 2001) reports on a study that examined the ability of resource authors to create acceptable – quality metadata in an organizational setting using manual evaluation by experts. Dushay & Hillman (2003) studies the issue of quality by pinpointing deficiencies that degrade it. In the same work, the use of a graphical tool to visualize the deficiencies in a repository level is also proposed. The issue of quality assurance is treated in (Barton, Currier & Hey, 2003; Guy, Powell & Day, 2004; Currier et al., 2004) and general principles and guidelines for the creation of metadata, in order to meet the functional requirements of the application in which they are used, are provided. In the context of quality assurance, (Hillman & Phipps, 2007) discusses the contribution of application profiles as a means for exposing and enforcing metadata quality.

A more systematic and organized view of metadata quality is achieved with the introduction of generic frameworks for the evaluation of quality. In (Moen, Stewart & McClure, 1997) a procedural framework for evaluating metadata records is introduced, using a set of 23 evaluation criteria. The framework discoursed in (Gasser & Stvilia, 2001) is based on concepts and ideas of the more generic field of information quality. It identifies 32 information quality parameters classified into 3 dimensions: intrinsic, relational/contextual and reputational. (Bruce & Hillman, 2004) elaborates on 7 characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility. Using (Bruce & Hillman, 2004) as a theoretical background, (Ochoa & Duval, 2006) attempts to operationalize the measurement of quality in a set of automatically calculated metrics for the 7 parameters. Similar efforts to provide metrics for metadata quality parameters can be found in (Hughes, 2004).

Focusing on the field of logic rules connecting metadata of related resources, the review of the related literature does not reveal any attempt to use such rules as a means to evaluate metadata quality. However, they have been used for automatic metadata generation. Duval & Hodgins (2004) points out that a resource's metadata may derive from the metadata of related resources. Hatala & Richards (2003) refers to resources being parts of a collection. In this case, it is possible that these resources share common values in their metadata elements. Although the resources in the collection and their metadata records are distinct, a value set for one metadata element in one resource can propagate itself to other resources of the collection. If the assembly is organized hierarchically, some of the values can be inherited from the ancestor nodes or aggregated from the child nodes. In other cases, the relations connecting the resources may not be such that the metadata value propagates as it is, but the value may be the result of a mathematical or logic expression of metadata of the related resources that either imposes a certain value, or restricts the range of values. Research efforts that use logic (or inference) rules for automatic metadata generation, either explicitly, or implicitly, are included in (Bourda, Doan & Kekhia, 2002; Brase, Painter & Nejdl, 2003; Doan & Bourda, 2005; Motelet, 2005; Margaritopoulos, Manitsaris & Mavridis, 2007). These efforts make use of the LOM metadata schema (IEEE, 2002).

Based on the background of the related work, this paper proceeds to define a new conceptual framework for metadata quality and a method for its assessment that exploits logic rules expressing interdependencies of the metadata.

3. The concept of metadata quality (the court metaphor)

The purpose of metadata is to provide adequate and correct information to their user so as to obtain a true picture of the content of a resource without having to access it. Any effort to approach metadata quality must always take this purpose into account. Metadata serve as the "mirror" of the resource, therefore their quality expresses the true representation of the resource and the absence of any distortion of its picture.

In order to approach the concept of quality, we can make use of a highly intuitive metaphor from a court of law. The metaphor defines a conceptual framework which can serve as a theoretical background to support the study of metadata quality. If we represent the resources of a repository with the facts of a case in court, the assessment of the quality of metadata is a process parallel to the evaluation of the descriptions of the facts of the case, as they are testified by the witnesses (with the assumption that for every fact there is only one witness). The (one and only) metadata record describing a resource in the repository is represented by the description of a fact by a witness (his/her testimony). The testimony of the witness comprises a set of single statements for every different aspect of the fact described. These statements represent the fields of the metadata record.

The issue of defining the quality of the metadata of a resource can be approached by using the abstract of the oath a witness takes in the court when he/she swears to "…tell the truth, the whole truth and nothing but the truth…" for the case he/she testifies. The quality of the testimony is assessed from its distance from the true fact ("truth" – correctness of the testimony), the inclusion of all the possible aspects of the fact ("whole truth" – completeness of the testimony) and the relation of the testimony with the case under examination ("nothing but the truth" – relevance of the testimony). The representation of the resources in a repository with the facts of a case in court and the metadata describing the resources with the witnesses' testimonies, leads to defining metadata quality as the resultant of their correctness, completeness and relevance.

The correctness of metadata refers to the intellectual distance separating them from the true representation of the resource being described. Correctness can be classified into two levels: The first, lower level concerns the requirement that the values of the metadata fields must obey the grammatical and syntactical rules of the language and the metadata standard or the application profile used. Missing letters, misspelled words, inconsistent formatting or representation of the same fields, fields containing inappropriate values according to the standard, are among the problems of this level. A metadata record must strictly follow the rules and guidelines of the standard or the application profile in order to be correct, just like a witness must be able to properly use the language to communicate in order to set his/her testimony fully understandable and, thus, allow the jury to form an opinion on its truthfulness. The second, higher level of correctness requires the semantical rightness of the values of the metadata fields, that is, the true representation of the reality and the absence of any deception. In court terms, this level refers to the truthfulness of the testimony. The first level of correctness concerns objective information, and for the purposes of this paper it is considered to be resolved, for example, by using any relevant validation parser. The second level of correctness is more subjective and it is the one that will be dealt with in this paper.

The completeness of metadata refers to their sufficiency to fully describe a resource. In essence, completeness measures the presence or absence of values in the metadata fields. In the court metaphor, completeness of a testimony refers to the adequate coverage of all the aspects of the fact described by a witness and to the provision of answers to all the questions he/she is asked. The choice of questions addressed to the witnesses is a task performed by the judge. The choice of the metadata fields – where values are to be filled in, in order for the record to be considered complete – is a matter of the requirements of any given application. Thus, in a given application context, other fields are considered as important and others are not. The application profile plays the role of the judge by selecting certain metadata fields to be accounted as mandatory or optional. A remarkable study on the completeness of a metadata record, focusing on educational resources (learning objects), is included in (Sicilia et al, 2005).

The relevance of the metadata of a resource has to do with its context of use. A metadata record of absolute correctness and full completeness may not be of quality if the (complete and correct) values of the metadata fields do not comply with the context of use. A testimony of a witness in the court, although complete and true, might be irrelevant with the case. This means that the context of a question asked to the witness is incompatible with the context of his/her answer to the question, because of possibly different perspectives. Relevance, as a component of quality, is highly subjective and may be confused with correctness, in the sense that faulty values might be due to either incorrectness, or irrelevance. However, the discriminating factor is the context. An incorrect value is faulty regardless of the context, while an irrelevant value is associated with a particular context. For example, a faulty value for the metadata field "Date of creation" of a resource is a matter of incorrectness (either syntactical – wrong format of the real date of creation, or semantical – a syntactically correct date different from the real one) regardless of any possible context. On the other hand, (although correct) values of the metadata field "Keyword" of a digital photo may be faulty due to irrelevance regarding a given context. If the digital photo has been indexed in a museum of photography, its keywords might be irrelevant when the photo is used in an image processing course. A way to reduce subjectivity and increase the relevance of the metadata is the use of vocabularies of values. A judge in the court restricts the witness's possible answers with the use of similar vocabularies ("…please answer with a yes or no…"). In this logic, a faulty value in a metadata field with a range of values out of a vocabulary will be, more possibly, attributed to incorrectness, rather than irrelevance.

The concept of quality is approached in the proposed conceptual framework by identifying the fundamental components and explicitly stating a solid definition which is domain and method independent. This definition targets the notion of metadata quality, directly. In a different sense, several of the related studies of metadata quality referenced above (Stvilia, 2001; Hillman, 2004; Moen, Stewart, & McClure, 1997) try to locate characteristics of metadata indicating quality or to detect deficiencies indicating its absence. Since no researcher claims to have found an exhaustive list of such characteristics, although this list is necessary to have quality, being not sufficient, it cannot guarantee its existence. Conversely, only if quality exists, all of the proposed characteristics are considered to be present. Such characteristics include parameters or dimensions of quality, like "accuracy", "precision", "naturalness", "informativeness", etc. Some other characteristics constitute signs and trails implying quality and not indicators assessing quality itself. For example, the parameter "provenance" (Bruce & Hillman, 2004) corresponds to the level of reliability and expertise of the metadata record creator. However, although the value of the creator of metadata is a good starting point to assume quality of his/her product, it cannot serve as a proof for quality, for the same reason a testimony of a witness cannot be considered to be true, only because of his/her high social acceptance and respectability. One could say that provenance assesses the probability of having quality in metadata. Other parameters in this category include "timeliness", "currency", "conformance to expectations", "volatility", "authority".

The conceptual framework for metadata quality presented in this section provides the necessary background to support methods and techniques for assessing quality. A method for quality assessment exploiting logic rules that correlate metadata fields and records will be introduced in the next section.

4. A Method for Metadata Quality Assessment Using Logic Rules

Keeping on the court metaphor, one can say that the verdict of the court for the case under examination is based on the assessment of the quality (i.e. correctness, completeness and relevance) of all the witnesses's testimonies. With the already stated assumption that for each fact there is one and only witness account (each resource in the repository is described by only one metadata record), a method for assessing the quality of a testimony is to check for the presence of inconsistencies; on the one hand to check for contradicting descriptions regarding the aspects of the fact and on the other hand to check for contradictions when the testimony is examined in comparison with other testimonies describing related facts. Any such contradictions violate implied logic rules and cause the testimonies to be considered unreliable. Compliance of a testimony with these rules classifies it as reliable.

In this sense, in order for a metadata record to be of quality, it has to comply with similar rules expressing logic dependencies, both among fields inside the record and among fields of records of related resources. A method to assess metadata quality is to check for the validity of logic rules expressing these dependencies.

4.1. Dependencies of Metadata Fields

In some cases, the fields of a metadata record are not completely independent from each other denoting intra-record dependencies. They present some sort of correlation, which is implicitly (if not explicitly) imposed by the specifications of the standard. The degree to which the values of correlated fields inside the record conform to the logic dictated by the relation between the fields is an indication of the record's quality. For example, the fields «1.7 General.Structure» and «1.8 General.Aggregation Level» of LOM are directly interdependent, as it is dictated by the LOM specification (IEEE, 2002), according to which "a learning object with Structure="atomic" will typically have AggregationLevel=1". The violation of this rule indicates degraded quality of the record.

Of course, the existence of relations between the fields of a metadata record indicates a "weakness" of the metadata schema, since, "…an efficient metadata system strives to have as nearly independent dimensions as possible…" (Wason & Wiley, 2001). However, the exclusion of such interdependences between the fields of a record is not always possible; hence, this fact is exploited for the evaluation of quality by examining the existence of certain combinations of values in the related fields inside the record (Ochoa & Duval 2006).

The dependencies of metadata fields are not restricted to fields inside a single record. They may concern fields of records of related resources denoting inter-record dependencies. Resources, related to each other with some kind of relation, create together a whole and therefore, it is possible that several of their metadata fields are influenced by each other. The influence of the values of the metadata fields is done on the basis of logic rules which constitute a set of validation principles that quality metadata fields must conform to. The definition of logic rules is an intellectual task, which has to take into account the semantics of the relations and the metadata. A methodology to create logic rules stemming from relations between metadata fields among records has been proposed in (Margaritopoulos, Manitsaris & Mavridis, 2007) for the purpose of metadata generation. The concepts and ideas presented in this work will serve as the starting point for defining logic rules to be used as validation rules for quality assessment of metadata, in the next subsection.

The core concept in the proposed methodology is the interrelated properties of the resources connected with a relation. These properties are called "connection features" and are specified on the basis of similarities or differences of the related resources. Connection features may be stated explicitly in the definition of the semantics of a relation. However, in other cases, connection features may be implied. For example, the definition of the semantics of the relation “IsVersionOf” of Dublin Core (DCMI Usage Board, 2008) clearly highlights the connection features “Format” and “Creator”, because the related resources have the same format and the same creator, whereas, one can presume the connection feature “Topic area” because different versions of a resource belong to the same topic area. Another example of a connection feature is “Intellectual content” deriving from the relation “IsFormatOf” of Dublin Core, since resources related with this relation have the same content. Apart from relations referring to semantic characteristics of the resources they connect, structural relations (part – whole relations) connecting the related resources are also included in the definition of connection features (“part” or “subset, “whole” or “superset” connection features).

The connection features, thought as properties of resources, can be mapped to certain metadata fields of the schema used for describing the resources. For example, the connection feature “Intellectual content” maps to metadata fields which express concepts and properties of learning objects exclusively influenced by their intellectual content. For the LOM standard, in these fields, «1.2 General.Title», «1.4 General.Description», «1.5 General.Keyword», «5.2 Educational.Learning resource type», «5.4 Educational.Semantic density», «5.6 Educational.Context», «5.7 Educational.Typical age range» are included.

The interrelation of the connection features of two resources (through the relation they are connected with) is translated into the interrelation of their respective metadata fields. These interrelations form a set of logic rules the violation of which indicates metadata records of degraded quality. An example of such rule for the metadata field “1.5 General.Keyword” of LOM can be derived from the connection feature “Intellectual content” of the relation “IsFormatOf”. “Intellectual content” feature can be mapped to this field because keywords are determined by the content of an object. The rule can be expressed as “learning objects that differ only in their format (they have the same content), must have the same keywords”.

4.2. Quality Assessment Rules

The logic rules, used for assessing quality of metadata utilizing related resources, can be distinguished into three major categories:

- *Rules of Inclusion*: the resource’s metadata field values must include the values of the same metadata field of records of related resources. Rules of inclusion apply only on metadata fields with cardinality greater than 1.
- *Rules of Imposition*: the resource’s metadata field values must be equal to the result of a mathematical or logic expression of metadata field values of the records of related resources (or of metadata field values of the same record, resulting from intra-record dependencies).
- *Rules of Restriction*: the range of a resource’s metadata field values is not the complete value space defined by the specification of the standard used, but a proper subset of it computed from the values of the same metadata field of records of related resources (or of another metadata field of the same record, resulting from intra-record dependencies).Values not belonging to this subset are prohibited.

In order to come up with a complete set of such rules, the semantics of relations connecting the resources and the semantics of metadata have to be taken into account. The rules influence the values of the metadata fields according to the category they belong to. It is obvious that the rules are metadata standard (or application profile) specific. For example, in the LOM standard a rule of inclusion dictates that the field “1.3 General.Language” of a learning object must include the values of the same metadata field of its parts (relation “HasPart”). Additionally, a rule of imposition imposes the value of the field “4.1 Technical.Format” of a learning object to be the same with the corresponding value of another learning object connected to the first one with the relation “IsVersionOf”. Moreover, a rule of restriction restricts the range of values of the field “5.7 Educational.Typical age range” of a learning object to be greater than the maximum typical age range of the objects it “Requires”.

A comprehensive list of rules is a matter every community of practice should deal with in the context of the application profile used. An important issue that remains open for consideration is the matter of conflicts. A conflict may come up when the value of the metadata field of a resource is influenced by two or more rules, according to the resource’s relations, yielding contradicting values. In this case a conflict policy must be defined.

5. Application of Quality Assessment Logic Rules

Given the definition of quality presented in Section 3, the logic rules deriving from relations among digital resources can be applied to their metadata in order to assess the fundamental components of their quality, i.e. their correctness, completeness and relevance.

For each metadata record in a repository, all the rules affecting the value of metadata fields are applied. Thus, according to whether a rule is valid or not, we infer the following:

- *Validity of a rule of inclusion:* If a rule of inclusion is valid, i.e. the metadata field of a resource under consideration includes values of corresponding metadata fields of its related resources, there is a clear indication of quality of all the involved fields. On the contrary, if such a rule does not hold, it is an indication either of reduced completeness of the field under examination, or of reduced correctness or relevance of its related fields. For example, as stated in the previous Section, in the LOM standard, a rule of inclusion dictates that the field “1.3 General.Language” of a learning object must include the values of the same metadata field of its parts (relation “HasPart”). Examining the validity of this rule for a learning object by comparing the values of its “1.3 General.Language” field against the value, e.g. “en”, of the same field of a learning object that is part of the first one, can lead to two results: If the rule holds, that is, the value “en” is included in the values of the field of the learning object under examination, then there is a clear indication of quality of the two involved fields. If the rule does not hold (the English language is not included in the values of the field of the learning object under examination), there are two cases: a) There is an indication of reduced completeness of the field of the first learning object. b) There is an indication of reduced correctness, either on the first, or on the second learning object (or on both). While in this example, concerning field “1.3 General.Language” of LOM, the problem of reduced quality in case b is, clearly, correctness, there might be situations where the faulty values derive from the context, so relevance might be the problematic component of quality.
- *Validity of a rule of imposition:* If a rule of imposition is valid, that is the resource’s metadata field values are equal to the result of the mathematical or logic expression of metadata field values of related resources suggested by the rule, then there is a clear indication of quality of all the involved fields. On the contrary, if the rule does not hold, there are two cases corresponding to this: a) The metadata field under examination does not have any value. The absence of value is a matter of reduced completeness. b) The metadata field under examination has a different value than the one dictated by the rule. In the case of a field with cardinality 1, the inequality of its value with the value dictated by the rule is an indication of absence of correctness (or reduced relevance) for the set of the involved fields. If the field under examination is of cardinality greater than 1, then the inequality of its (multiple) value with the value dictated by the rule, implies either completeness or correctness – relevance deficiencies (or both) for the set of the involved fields, depending on the relation between the set of values of this field and the set of values dictated by the rule. For example, a rule of imposition in the LOM standard, dictates that the value of the field “5.11 Educational.Language” of a learning object must be equal to the value of a learning object related to the first one with the relation

“IsRequiredBy”, in the sense that if a learning object is required by another one, then the human language used by the typical intended user of this object will be the same with the corresponding language of the object that requires it. Considering several combination of values, we have: If learning object x “IsRequiredBy” learning object y and “5.11” of x = “en”, while “5.11” of y = “en”, then the rule holds, and there is a clear indication of quality of the two involved fields. If “5.11” of x does not have any value, while “5.11” of y = “en”, then the rule does not hold, and there is indication of reduced completeness. If “5.11” of x = “en”, while “5.11” of y = “en”, “fr”, then the rule does not hold and there is indication of reduced completeness, as well. If “5.11” of x = “en, “fr”, while “5.11” of y = “en”, “it”, then the rule does not hold. This situation might imply either a problem of reduced correctness (“fr” has been mistakenly taken for “it”), or a problem of both correctness and completeness (“fr” has by error been included in the values of “5.11” of x, while at the same time “it” has been omitted from the set of the values). In the last case, the indicated quality problems do not concern only learning object x, but both related objects as a pair, since the quality of y has not been taken for granted.

- *Validity of a rule of restriction:* The validity of a rule of restriction, that is, the presence of a value of the metadata field under examination within the restricted range dictated by the rule, is an indication of quality. On the contrary, if the value of this field is outside the dictated range, it is a case of absence of correctness for the involved fields. For example, a rule of restriction in the LOM standard, dictates that the value of the field “1.8 General.Aggregation level” of a learning object must be less than the minimum aggregation level of its parts (the learning object is related with its parts with the relation “IsPartOf”). If such is the case, then the validity of the rule indicates quality of the involved objects. If the value of “1.8” of the learning object is not less than the minimum aggregation level of its parts, then the rule does not hold and there is an indication that the values of the involved fields are not correct.

The quality problems located by examining the validity of logic rules provide valuable hints to the administrators of the metadata repository. Although the method cannot locate the problematic component of quality, exactly on a single record or element, it restricts the field of interest and focuses on a reduced set of resources with degraded quality. This is evident, since the conclusions one can draw by examining the validity of the rules concern more than one (related) fields, where no field is considered to be of high quality in advance. In the general case, where no such assumptions deriving from the context of use or the specific application are made, the set of the related fields with problematic quality is the limit of the quality assessment’s “granularity”. However, this method combined with other methods of metadata quality assessment can be of valuable contribution. For example, metrics referenced in Section 3, or manual inspection by experts can be applied to the set of the fields not following a certain rule, in order to pinpoint the problematic ones. This is much more feasible and efficient compared to the usage of these methods over the whole repository.

The logic rules, which in this paper are proposed to be used as a means for quality assessment of the metadata, can also be used to enhance quality when chosen to be applied and modify the values of the involved fields. Used as metadata generation rules, they can increase completeness by populating empty fields, as well as correctness or relevance by replacing faulty values. Especially, the increase of relevance can be considered as a method to preserve the context in the metadata records, in cases where the records are created by various indexers with diverse backgrounds. Of course, all these benefits are a result of well established application policies on the fields to be considered of high quality as reference.

6. Conclusion and Future Work

In this paper, a new framework for conceptualizing metadata quality was defined using analogies from common knowledge and experience. The framework was inspired from entities and procedures involved in a court of law and aims at setting a solid, simplified theoretical background by defining the fundamental components of metadata quality, namely: correctness, completeness and relevance. Then, metadata quality assessment is performed by assessing these three components. Hence, the paper proposes a method for assessing metadata quality by exploiting structural and semantic relations among digital resources in a repository. Such relations create logic rules connecting the metadata of the related resources. Examining the validity of the rules serves as a means to conduct quality control on the metadata of the involved resources.

The conclusions deriving from this process can form the basis for a metric system to measure the components of metadata quality. Possible factors to be taken into account in the design of the metric system might be the number of non-valid rules at record or repository level, the number of the involved fields in a rule, the number of faulty or missing values in a field, the number of the resources participating in a problematic set, the number of problematic sets a single resource participates in, and so on. The design of such metrics is a step forward following this work. The method proposed in this paper can be combined with other metadata quality assessment methods and techniques in an integrated quality assurance system for the metadata of a digital repository.

References

Barton, Jane, Sarah Currier, and Jessie M. N. Hey. (2003). Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. *Proceedings of Dublin Core Conference 2003: Supporting Communities of Discourse and Practice - Metadata Research and Applications, 2003,* (pp. 39-48).

Bourda, Yolaine, Bichlien Doan, and Walid Kekhia. (2002). A semi-automatic tool for the indexation of learning objects. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2002,* (pp. 190-191).

Brase, Jan, Mark Painter, and Wolfgang Nejdl. (2003). Completion Axioms for learning object metadata - Towards a formal description of LOM. *3rd IEEE International Conference on Advanced Learning Technologies (ICALT 2003).*

Bruce, Thomas R., and Diane I. Hillmann. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In Dianne. I. Hillmann and Elaine L. Westbrooks (Eds.), *Metadata in practice,* (pp. 238-256). Chicago: ALA.

Bui, Yen, and Jung-ran Park (2006). An assessment of metadata quality: A case study of the national science digital library metadata repository. In Haidar Moukdad (ed.), *CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation* from http://www.cais-acsi.ca/proceedings/2006/bui_2006.pdf.

Currier, Sarah, Jane Barton, Rónán O'Beirne, and Ben Ryan. (2004). Quality assurance for digital learning object repositories: Issues for the metadata creation process. *ALT-J, Research in Learning Technology, 12*(1), 5-20.

DCMI Usage Board. (2008). *DCMI Metadata Terms.* Retrieved March 18, 2008, from http://dublincore.org/documents/dcmi-terms/.

Doan, Bich-lien, and Yolaine Bourda. (2005). Defining several ontologies to enhance the expressive power of queries. *Proceedings on Interoperability of web-based Educational Systems, WWW'05 conference, Chiba, Japan.*

Dushay, Naomi., and Diane I. Hillmann. (2003). Analyzing metadata for effective use and re-use. *DCMI Metadata Conference and Workshop, Seattle, USA.*

Duval, Erik, and Wayne Hodgins. (2004). Metadata matters. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004,* (pp. 11-14).

Friesen, Norm. (2004). *International LOM Survey: Report (Draft).* Retrieved March 18, 2008, from http://dlist.sir.arizona.edu/403/01/LOM%5FSurvey%5FReport2.doc.

Gasser, Les, and Besiki Stvilia. (2001). A new framework for information quality. *Technical report, ISRN UIUCLIS–2001/1+AMAS, 2001.*

Greenberg, Jane, Maria Cristina Pattuelli, Bijan Parsia, and W. Davenport Robertson. (2001). Author-generated Dublin Core metadata for web resources: A baseline study in an organization. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2001,* (pp. 38–46). National Institute of Informatics.

Guy, Marieke, Andy Powell, and Michael Day. (2004). Improving the quality of metadata in Eprint archives. *Ariadne 38.*

Hatala, Marek, and Griff Richards. (2003). Value-added metatagging: Ontology and rule based methods for smarter metadata. In Michael Schroeder and Gerd Wagner (Eds.), *RuleML 2003,* (pp. 65–80).

Hillmann, Diane. I., and Jon Phipps. (2007). Application profiles: Exposing and enforcing metadata quality. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007,* (pp. 52-62).

Hughes, Baden. (2004). Metadata quality evaluation: Experience from the open language archives community. *Digital Libraries: International Collaboration and Cross-Fertilization,* (320–329).

IEEE. 1484.12.1 (2002). Draft Standard for Learning Object Metadata*. Learning Technology Standards Committee of the IEEE*. Retrieved 18 March, 2008, from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf.

Margaritopoulos, Merkourios, Athanasios Manitsaris, and Ioannis Mavridis. (2007). On the Identification of Inference Rules for Automatic Metadata Generation. *Proceedings of the 2nd International Conference on Metadata and Semantics Research (CD-ROM), 2007.* Ionian Academy.

Moen, William E., Erin L. Stewart, and Charles L. McClure. (1997). Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS). *Proceedings of the Advances in Digital Libraries Conference, 1998,* (p. 246). IEEE Computer Society

Motelet, Olivier. (2005). Relation-based heuristic diffusion framework for LOM generation. *Proceedings of 12th International Conference on Artificial Intelligence in Education AIED 200.* Amsterdam, Holland: Young Researcher Track.

Najjar, Jehad, Stefaan Ternier, and Erik Duval. (2004). User behavior in learning object repositories: An empirical analysis. *Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, AACE, 2004,* (pp. 4373–4379).

Ochoa, Xavier, and Erik Duval. (2006). Quality Metrics for Learning Object Metadata. In Elain Pearson and Paul Bohman (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006,* (pp. 1004-1011). Chesapeake, VA: AACE.

Sicilia, Miguel-Angel, Elena García-Barriocanal, Carmen Pagés, José-Javier Martínez, and José-María Gutiérrez. (2005). Complete metadata records in learning object repositories: Some evidence and requirements. *International Journal of Learning Technology, 1*(4), 411-424.

Wason, Thomas D., and David Wiley. (2001). Structured Metadata Spaces. In Jane Greenberg (ed.), *Metadata and Organizing Educational Resources on the Internet,* (pp. 263-277). New York: Haworth Press.

Full Papers

Session 5:

Tagging and Metadata for Social Networking

Semantic Relation Extraction from Socially-Generated Tags: A Methodology for Metadata Generation

Miao Chen	Xiaozhong Liu	Jian Qin
Syracuse University, USA	Syracuse University, USA	Syracuse University, USA
mchen14@syr.edu	xliu12@syr.edu	jqin@syr.edu

Abstract

The growing predominance of social semantics in the form of tagging presents the metadata community with both opportunities and challenges as for leveraging this new form of information content representation and for retrieval. One key challenge is the absence of contextual information associated with these tags. This paper presents an experiment working with Flickr tags as an example of utilizing social semantics sources for enriching subject metadata. The procedure included four steps: 1) Collecting a sample of Flickr tags, 2) Calculating co-occurrences between tags through mutual information, 3) Tracing contextual information of tag pairs via Google search results, 4) Applying natural language processing and machine learning techniques to extract semantic relations between tags. The experiment helped us to build a context sentence collection from the Google search results, which was then processed by natural language processing and machine learning algorithms. This new approach achieved a reasonably good rate of accuracy in assigning semantic relations to tag pairs. This paper also explores the implications of this approach for using social semantics to enrich subject metadata.

Keywords: relation extraction; tags; search engine; social semantics; metadata

1. Introduction

The recent social tagging movement has generated abundant semantic resources for representing the content of information objects. Unlike traditional subject indexing performed by trained librarians, the socially-generated semantic tags are created by users who want to assign tags to the information objects of their interest. While these tags are sometimes erroneous and ill-constructed (Guy & Tonkin, 2006; Mathes, 2004; Michlmayr, 2002) this newfound wealth of social semantics has become a mining ground for discovering and understanding social networks and cultural taste (Liu et al., 2006; Mika, 2005), ontological structures (Schmitz, 2006), and various semantic relationships among the tags (Rattenbury et al., 2007).

Subject representation as one important area in metadata description may employ social semantics or controlled semantics. The two types of semantics can benefit each other in a profound way as Qin has discussed (2008). On the one hand, social semantics as empirical knowledge can contribute to controlled semantics through testing it and thus learning from it. On the other hand, social semantics provides a valuable source of empirically-derived knowledge to enrich and validate controlled semantics (Qin, 2008). We are facing, however, a number of challenges in accomplishing these goals. One such challenge is the methodology.

Tag mining methodology includes a wide variety of techniques and algorithms used to acquire, preprocess, parse, and analyze tag data. Before tag data becomes usable for mining tasks, it needs a series of linguistic, syntactic, and semantic processing. This processing is often computationally intensive and requires linguistic and semantic sources to be adapted to the mining techniques and tasks. Research on mining social tags to discover semantic patterns and relationships has applied machine learning, clustering, natural language processing, and other techniques (all which are reviewed in the next section).

A major weakness (among other flaws) of user-generated tags is the lack of semantic relations between terms, which are often represented in controlled semantics as broader, narrower, and related terms; or, in ontologies as relations between classes of concepts and instances. While it is impractical to expect users to categorize tags or provide semantic relations in the same way as librarians do for controlled semantics, it is possible to extract semantic relations using computational methodologies. The study reported in this paper is an attempt to address this methodology challenge. By using Flickr's tags as the source, we applied natural language processing (NLP) and machine learning techniques, in addition to Google search results, to the processing and analysis of Flickr tag data. The goal of this research has been twofold: 1) to experiment with an approach employing NLP and machine learning techniques combined with Web search results to provide the context of tags for extracting semantic relations from social semantics; and 2) to evaluate the effectiveness of this methodology. The long-term goal has been to develop effective methods for meshing up social and controlled semantics that can be used for subject metadata representation of digital objects and resources.

2. Literature Review

Semantic relations between concepts or entities exist in textual documents, keywords or key phrases, and tags generated in social tagging systems. Relation extraction refers to the identification and assignment of relations between concepts or entities. Automatic extraction of semantic relations has a wide range of applications in knowledge organization and information retrieval. Relation extraction can explore relationships that are implicit to underlying data and then add new knowledge to the different domains.

Previous studies have focused on relation extraction between entities from (document) textual resources. In traditional relation extraction, the sources of entities usually come from terms in unstructured documents such as Web pages or structured documents such as relational databases. A wide variety of data sources have been used in relation extraction research, e.g., Web pages (Brin, 1998), corpus (Bunescu & Mooney, 2007), and socially generated Wikipedia articles (Nguyen et al., 2007). The semantic and linguistic sources for exploring relations can be a corpus containing the context of entities, and this context information can serve as the basis of relation assignment.

No matter which data sources are utilized in relation extraction, it is necessary to meet three requirements: 1) a collection of data (entity) sources from which semantic relations will be extracted, 2) a semantic or linguistic source in which the context for relations is provided, and 3) algorithms for automatic execution of processing operations. How well a relation extractor performs is determined mainly by the context sources and algorithms. Context containing entities or concepts play a critical role in ensuring the precision of text relation extraction since this provides the source in which covert relations may inhabit.

While text relation extraction relies heavily on the context, current research on tag relation extraction rarely includes context information in the procedure. Tag relations are extracted by applying statistical methods to derive relations from tag co-occurrences, similarity computations, and usage distribution. Examples of these types of studies include a hierarchical taxonomy built from the Deli.cio.us and CiteULike tags by using cosine similarity of tag vectors (Heymann & Garcia-Molina, 2006), and an ontology generated from Flickr tags using statistical methods (Schmitz, 2006) that in turn was based on Sanderson and Crofts' (1999) model for the co-occurrences of tags. For each frequently co-occurring pair of tags, the model was applied to determine whether or not there was a hierarchical relation between them. Subsequently, a hierarchical structure of tags became an ontology. Rattenbury et al. (2007) presented an approach of identifying event and place tags from Flickr. The assignment of tags' semantic types was learned from patterns of temporal and spatial tag usages employing statistical methods. When contrasted with the three requirements of text relation extraction, it becomes apparent that the second requirement for context is missing from these tag relation extraction experiments.

Although the abovementioned methods have achieved varying levels of success, the absence of context information in these methods limits not only the accuracy of processing but also the scalability of automatic relation extraction. Our strategy in addressing this limitation was to add context to tags. By tracing tags to the context where they might have originally appeared or commonly been used, we could explore the context that would assist us to extract accurate and reliable relations. The methodology we employed involved using external document resources that have sentences containing the tags in the source data. Relations were then extracted from these documents and assigned to related tags.

Extracting semantic relations from documents is not a new area of research; in fact, a large number of studies on extracting relations from text (including Web pages) and corpus have been published in the last two decades. Relation extraction generally involves two primary parts: 1) the natural language processing (NLP) part, and 2) the machine learning part. NLP techniques are applied in order to identify entities and relation indicators from texts. Machine learning algorithms are implemented to learn features of relations, and assign relations to entities whose relations are not yet known. Text relation extraction also involves entity extraction for identifying entities or concepts (Brin, 1998; Iria & Ciravegna, 2005; Nguyen et al., 2007; Roth & Yih, 2002).

The NLP part of relation extraction is a process by which the text processing may be performed at different levels throughout different stages using either shallow processing or deep processing. Shallow processing involves sentence segmentation, tokenization, part of speech (POS) tagging, and chunking of the text being processed—which is used to identify phrases and chunks (Bunescu & Mooney, 2007). An example is the study by Roth and Yih (2002), where shallow parsing was used to segment sentences and to identify entities and relations. Deep processing builds a parsing (or dependency) tree by identifying the shortest-path dependency of language components in sentences. This NLP technique is useful when the context of pairs of entities needs to be processed. In such cases, the words located before, between, and after these entities are used directly as vectors for matching patterns of relations (Agichtein & Gravano, 2000). The question of whether to use a shallow or deep level of text processing is determined by the design of experiment(s) and algorithm of machine learning. If shallow processing is sufficient, then there is no need to use deep processing (Zelenko et al., 2003).

Machine learning performs a different role in relation extraction. As computer algorithms, machine learning is dependent upon features (variables) representing objects as the input into learning models. The features needed for machine learning may be entity types, words, phrases, part of speech, chunks, tags, etc. from the context sentence or sentence part (Bunescu & Mooney, 2007; Culotta & Sorensen, 2004). From samples (context containing pairs of entities) whose features and relation types are already known, machine learning generates patterns of different relations based on features. Subsequently, the generated patterns can be applied to new contexts with unknown relations and derive meaningful relations. Commonly used machine learning models include the support vector machine (SVM) (Bunescu & Mooney, 2007; Culotta & Sorensen, 2004; Zelenko et al., 2003), clustering (Agichtein & Gravano, 2000), undirected graphical models (Culotta et al, 2006), and decision tree (Nahm & Mooney, 2000).

A review of previous studies shows that past research in tag relation extraction has rarely used contextual sources for relation recognition and has seldom utilized techniques from text relation extraction. Tag relation extraction as a special case of relation extraction does not need entity extraction (because tags are not sentence-based documents) as does regular text relation extraction. To leverage the social semantics power for subject metadata description, we are faced with challenges brought about by the lack of context information in tag sources. Solving this problem is a critical first step to successfully deploying social semantics in subject metadata description. We will introduce the details of the proposed methodology for improving tag relation extraction in Section 3, the experiment using our methodology in Section 4, the results and performance in Section 5, and discussion of the results and conclusions in Section 6.

3. Methodology

In this section, we introduce our methodology in detail and explain the process of extracting relations between Flickr tags. Two sources are critical in this process: the source of entities and the source of context. Since entities have already been "extracted" by taggers, we instead focus on obtaining the context of the tags in our sample data by using search results from a general search engine.

As mentioned above, a major challenge in extracting tag relations is the lack of context information for the tags, which makes them insufficient and difficult for the relation extraction task. An example is a photo in Flickr that has been assigned four tags: *Shangrila* (a remote area in southwest China), *Mountain*, *Yunnan* (one of the provinces in China), and *River*, as shown in FIG. 1. Since the context is the photo itself and separated from the tags in the search system (i.e., image-based search is still not available in most search systems), the four tags could have a wide variety of contexts for interpretation when separated from the photo they describe.

FIG. 1. A photo in Flickr with four tags: Shangrila, Mountain, Yunnan, River.

From the perspective of relation extraction, photos do not provide sufficient context for tags and the relations between tags are not explicit. Due to technological limitations, it is difficult to process images in order to acquire semantics. Compounded by the technology limitation is the tagging practice that does not label any relations between the tags, e.g., relation "Shangrila is located in Yunnan" is information separate from either the photo or the tags. Acquiring tag relations without context information is analogous to a simple keyword search on the Web—the precision and recall can be very problematic. These predicaments led us to seek external text resources such as search engine results as a solution to obtaining the context of tags. A unique advantage of using tags to extract relations is that the entities are already "extracted" by human taggers and so the final error rate can be reduced by avoiding the errors that are propagated by the entity extraction process.

3.1. Identification of Problem

Given a set of tags from social tagging Web sites, our task was to discover relations between any two tags that frequently co-occurred. We defined our tag set as (Tag_1, Tag_2, Tag_3, Tag_4, …, Tag_n) ($n \in N$) and used statistical techniques to identify frequently co-occurring pairs of tags in the tag set. The selected tag pairs were then deposited in a new set called "tag pairs." A tag pair may be represented as *pair* (Tag_x, Tag_y), where Tag_x and Tag_y meet the requirement that both tags frequently occur together. Once the set of tag pairs was constructed, the next step was to identify the relation between Tag_x and Tag_y for each pair in the set.

To precisely and effectively identify relations between pairs of tags, the critical component is the context of tag occurrence. We determined that an effective method was to put the tag pairs back into context by employing results from a general search engine, and then applied natural language processing and machine learning techniques to extract relations from that context. The task at this stage included finding the context for tag pairs and building a classifier for relation assignment. For a tag pair (Tag_x, Tag_y), the relation was defined as R_{xy}, representing a single type of relation between Tag_x and Tag_y.

3.2. Assumptions

We made two assumptions regarding to tag relations. First; if two tags frequently co-occurred, there ought to be some type of relation between them or else they would not be frequently tagged together by users. A high frequency of co-occurrences is not coincidental; rather, it underscores the possibility of some connection between the tags. For example, since "San Francisco" co-occurred frequently with "bay area," we assumed that there was a strong possibility that a relation existed between the two tags. From our knowledge, the two have a relation that San Francisco "*is located in*" the bay area.

The second assumption: there is only one single relation between the two tags in a pair. It is possible that the two have more than one relation, e.g., San Francisco can be "*located in*" the bay area (San Francisco Bay) or San Francisco can be "*located in the north part of*" the bay area. When our human coders were assigning relations to tag pairs for the training data set, they assigned the most general and higher-level relations to the tag pairs. Using the San Francisco example, the relation is "*located in*" (since it is a higher level relation) that includes the instance of "*in the north part of.*" This assumption facilitated the extraction of more features for the learning model.

3.3. Selection of Tag Pairs

We downloaded 28,737 photos with 289,216 accompanying tags about landscape from Flickr– which contained 21,443 unique tags. These all co-occurred with (and are about) the tag "landscape." We used an index of mutual information to find pairs of tags that frequently co-occurred. The mutual information (MI) index between any two tags was calculated based on the co-occurrence between two tags, which was also used to describe and normalize the co-occurrences between two tags. The MI index represents the degree of relatedness in candidate tag pairs (Tag_x, Tag_y), i.e., the higher the MI scores, the more closely related the two tags are. The MI index was calculated by using the well-established formula below (Shannon, 1948):

$$MI(Tag_X, Tag_Y) = P(Tag_X, Tag_Y) \cdot \log \frac{P(Tag_X, Tag_Y)}{P(Tag_X) \cdot P(Tag_Y)} \qquad \text{[EQ. 1]}$$

For the tag set (Tag_1, Tag_2, Tag_3, Tag_4, …, Tag_n), we calculated the MI scores for any two tags, which resulted in an *n* x *n* matrix. Tag pairs with low MI scores were removed from the matrix and the remaining high MI score pairs were retained.

3.4. Relation Extraction

Having prepared tag pairs for relation extraction, the next step was to identify the context for tags and generate machine learning models for relation extraction. This process involved 1) entering a tag pair in a search engine query, 2) obtain search results, and then process the search results with NLP tools, 3) establish learning relation patterns from samples with known relation types, and then 4) derive candidate relations for tag pairs. FIG. 2 demonstrates the process of tag relation extraction.

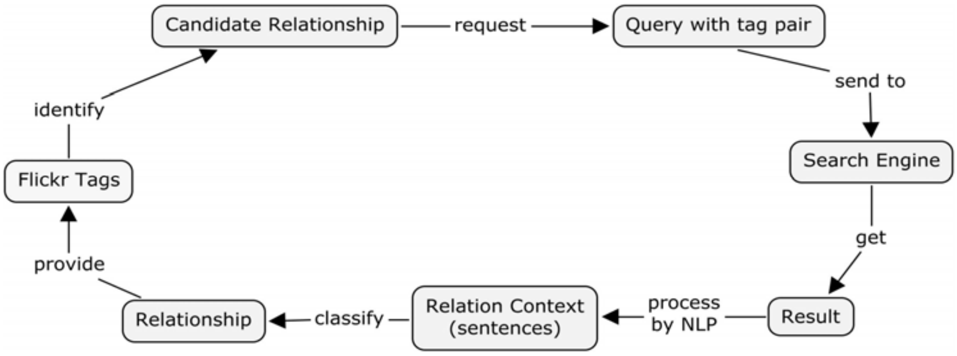


FIG. 2. Tag relation extraction process.

The tag pair query in the general search engine returned a list of results with a title and brief description for each resource in the result set. We assumed that such search results would provide the context for the tags if both tags in a pair appeared together in the results that were highly relevant to the query. If sentences from the search results contained both Tag_x and Tag_y, the sentences were then considered as the context of relation R_{xy} between Tag_x and Tag_y. Although not every returned sentence contained both Tag_x and Tag_y, the only ones needed contained both tags to use as context. Sentences meeting this criterion were selected for the context sentence collection.

Sentences in the context sentence collection were then parsed and chunked using NLP techniques. We applied the deep processing technique because it enabled us to learn more about the features of the context. The NLP processing returned a parsed sentence with part-of-speech tags of words and chunking tags of phrases. For example, a sentence "The largest city in the Sonoran Dessert is Phoenix, Arizona" is parsed into a tree (shown in FIG. 3).

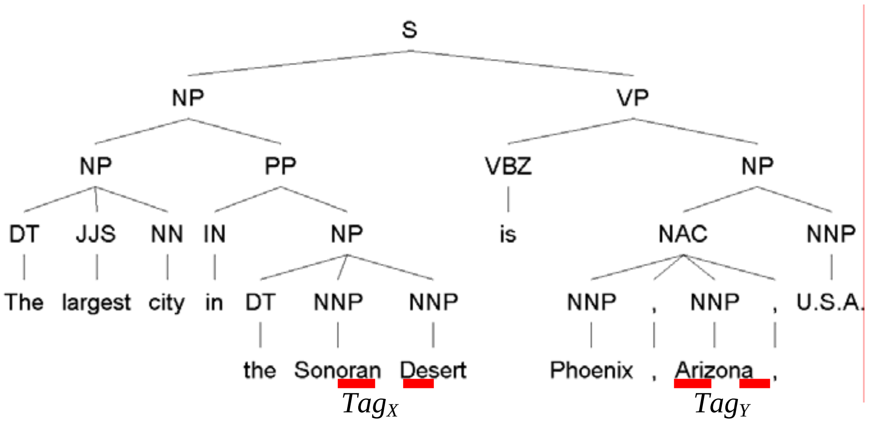


FIG. 3. Parsed tree of a sentence.

As already mentioned, a list of statistical and semantic features can be extracted from natural language processing results. Following Bunescu and Mooney (2007), we used the features of context before Tag_X, between Tag_X & Tag_Y, as well as after Tag_Y. The types of features we chose included: word (the word was processed by a Porter stemming algorithm for stemming), part of speech (e.g. verb, noun, and preposition), chunking, dependency subtree, and the distance between source feature and target feature. In the example sentence from FIG. 3, Tag_X is Sonoran Desert, Tag_Y is Arizona, and the goal is to find the relation between the two tags. The features scrutinized included: (Verb, between_Tag_X_Tag_Y), (verb, is) (DT, before_Tag_X, distance-1), (Tag_X, exist_in_NP), (Tag_Y, exist_inVP), (Tag_X, Tag_Y, lowest_common_father_S), and so forth.

Verb, between_Tag_X_Tag_Y) means that the verb between Tag_X and Tag_Y was taken as one feature, and other listed features can be similarly interpreted.

Once the relation between sample tag pairs was known, features of the tag context were input into machine learning algorithms to generate patterns of different relations. The machine learning algorithm applied the decision tree technique and features were selected to build a classifier for relation extraction. When a new tag pair was identified, the processing went through the above steps for identifying the context through search engine results and natural language processing. The resultant features were then entered into the classifier which later returned the relation type for the tag pair.

4. Experiment

As described in Section 3.3, the dataset contained 289,216 tags. The criterion for including a tag in the dataset was that if a tag appeared together with "landscape" for one or multiple photos, this tag would be included in the tag set. This selection process yielded 21,443 unique tags in the landscape domain.

Each tag pair in the tag set was then computed to generate a matrix of mutual information scores. Tags appearing less than 5 times in the tag set were deleted in order to reduce computation cost. After ranking the mutual information scores (from high to low) in tag pairs, the first 3,000 tag pairs were selected to form the tag pair set. Some example pairs are shown in the following table:

TABLE 1. Examples of tag pair's mutual information.

Tag_X	Tag_Y	*Mutual Info*
bay area	golden gate bridge	0.071521409
Backpacker magazine	CDT PROJECT	0.05926981
beach	ocean	0.058470874
beach	Florida	0.01961479
Beach Houses	vacation	0.015982523
aguila	snake	0.011943919
Acadia	Acadia National Park	0.011012993

As with the examples above, if Tag_X and Tag_Y have a high mutual information score, we can assume that there exists a strong relationship between Tag_X and Tag_Y, then marked as "Tag_X, candidate_relationship_?, Tag_Y." We used the following algorithm to identify candidate relationships:

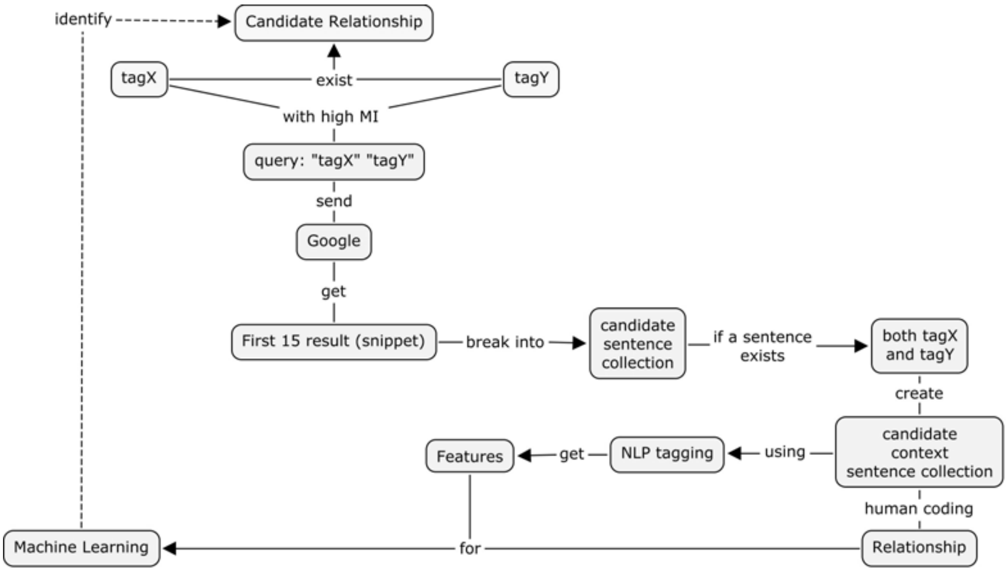


FIG. 4. Process of identifying candidate relationships between tags.

We chose Google as the general search engine whose results would provide context for relation extraction. Google has an API that provides snippet of retrieved sentences. By sending a tag pair (*Tag_X*, *Tag_Y*) as a query ("*Tag_X*" "*Tag_Y*") to the API, we received a snippet of each result. We used quotation marks in the query because both *Tag_X* and *Tag_Y* might be phrases rather than single words, and quotation marks ensure a more exact match for target phrases.

Snippets of the first 15 search results for tag pairs were exported and processed by a sentence boundary tool to identify sentences and put them into candidate sentence collection. The program tested each sentence to see whether or not it contained both *Tag_X* and *Tag_Y*. If so, this sentence would be included in the candidate context sentence collection. There was often more than one sentence in the snippet satisfying this requirement for contextual information.

Two human coders manually marked relations for a small portion of the tag pairs in the sentence context collection. Since one tag pair might have more than one context sentence—while only one relation type can be assigned to a tag pair regardless how many context sentences it might have—the most general and high-level relation was assigned to the tag pair. The manual coding produced eight types of relations: 1) *is-a-measure-of*, 2) *is-located-in*, 3) *induces*, 4) *is-induced-by*, 5) *is-style-of*, 6) *is-of*, 7) *is-for*, and 8) *is-a-method-of*. Examples of the relation between tag pairs and context sentences are presented in the following table:

TABLE 2. Human coded relations and context sentences.

tagX	**relation**	***tagY***	**Context Sentence**
2-deoxy-d-glucose	induces	effect	Effect of 2-deoxy-D-glucose on cell fusion induced by Newcastle disease and herpes simplex viruses.
2-DG	induces	effect	Effect of peripheral 2-DG on opioid and neuropeptide Y gene expression.
Action	is-induced-by	anticonvulsant	Pharmacokinetic modeling of the anticonvulsant action of phenobarbital in rats. J Dingemanse, JB van Bree and M Danhof.

tagX	**relation**	***tagY***	**Context Sentence**
Action	is-induced-by	epilepsy	From Epilepsy Action, the UK's leading epilepsy charity.
Acadia	is-located-in	maine	Use this vacation and travel guide to the Downeast and Acadia region of Maine to plan your vacation, business trip or just for fun.
Alabama	is-located-in	America	The Alabama Location Map indicates the exact geographical position of the states of the United States of America.
America	is-located-in	san francisco bay area	Boy Scouts of America, San Francisco Bay Area Council • 1001 Davis Street, San Leandro, CA 94577-1514, (510) 577-9000.

In the preliminary experiment, we chose three relation classes (for 121 cases) for machine learning tasks from human coded relations: *"induces,"* "*is-induced-by*," and "*is-located-in."* Among the 121 sentences, part of them were used as the training set for feature extraction and model building, and the remainder were used for evaluation.

We applied the Stanford parser for parsing and chunking in the NLP phase. This step was to prepare for the machine learning part. The parsing of context sentences generated candidate features for machine learning, and when combined with features and relation labels we were able to then conduct training to derive a classifier for relations. A decision tree was the algorithm for selecting features and generating patterns for different types of relations. The resultant classifier was then ready for accepting new context for tag pairs and outputting relations.

Finally, we examined the methodology by sending new tag pairs to the trained model. We withheld the other context sentences as a testing set, and input the sentences to the NLP processor and classifier accordingly. The classifier returned the relation of each tag pair as an automatic relation extraction result. Since we had the human coded results, we compared them with the machine learning results and evaluated the performance.

5. Results and Analysis

Our preliminary experiment extracted 2401 unique features from 121 context sentences. We used a ten-fold cross-validation to evaluate the result (Table 3). The evaluation result of our method displayed in Table 3 shows an 83.72% rate of correct classification / tag relation instances. While the sampling size of tag data and the number of human coded relations could not be as large as we would have liked, this approach appears to be a promising methodology. The introduction of external sources allows for objectively identifying contextual information for context-less tag data and thereby improving the accuracy and reliability of relation extraction.

TABLE 3. Evaluation result of the preliminary experiment.

	is-located-in	*is-induced-by*	*induces*
is-located-in	90	2	1
is-induced-by	10	12	3
induces	1	4	6
Correctly Classified Instances 108 83.72 %			
Incorrectly Classified Instances 21 16.28 %			

This result also suggests that using external sources for context information can help detect data anomalies in the tag pairs that have a high MI score. We discovered from our experiment that a high MI score did not necessarily mean that Tag_X and Tag_Y always had direct semantic relations. Some tag pairs did not appear in any sentence in Google search results and no context was found containing the tags. For example, the two tags Tag_X = "all rights reserved" and Tag_Y = "Canon EOS 350" had a high MI score, but neither of these two tags appeared together in Google search results; this suggests that no context sentence existed for the two tags. We are unsure at present how the two tags might be related, but it is possible that they are indirectly related. If, for example, they are both related to a third tag in a meaningful way, then they could be related to each other statistically but not semantically. Consequently, the two tags were semantically unrelated and the pair was removed from the tag pair collection to ensure the meaningfulness of tags and their relations.

We also discovered that NLP algorithms can provide flexible and powerful features for relation identification. For instance, a syntax level feature can be helpful for identifying the "*is-located-in*" class in an example pattern such as Tag_X, Tag_Y, Zip Code or Tag_X prep Tag_Y (*prep* could be "in," "with," or "by"), where Tag_X could be a city name and Tag_Y a state name. The NLP algorithms then can be expanded and explored with more semantic feature types and other machine learning algorithms.

6. Conclusion

Tags are a special type of subject metadata as well as a rich, powerful vocabulary source. Extracting relations between tags is the first step toward automatic subject metadata creation. An important contribution of this study was the introduction of external resources as a solution to the problem of context-less tag data. Through combining NLP and machine learning techniques we developed a set of algorithms and procedures for automatically processing the external resources, using the output to provide more objective, reliable context information for tag relation extraction.

The methodology developed in this study can be applied to larger-scale research in the future as well as in research fields beyond tag relation extraction. For example, the processing and categorization of unstructured text can benefit from this methodology, as can automatic construction of an ontology and controlled vocabulary, as well as automatic mapping between tags and controlled vocabularies.

The results of our approach are encouraging for tag relation extraction. We plan to improve the classifier by collecting more relation types and human-coded examples for future experiments, and eventually utilize the relations extracted to enhance subject metadata descriptions.

References

Agichtein, Eugene, and Luis Gravano. (2000). Snowball: Extracting relations from large plain-text collections. In Kenneth M. Anderson, et al. (Ed.), *Proceedings of the 5th ACM Conference on Digital Libraries,* (pp. 85-94). New York: Association for Computing Machinery.

Brin, Sergey. (1998). Extracting patterns and relations from the World Wide Web. In Paolo Atzeni et al. (Ed.), *Selected Papers from the International Workshop on the World Wide Web and Databases,* (pp. 172-183). London: Springer.

Bunescu, Razvan C., and Raymond J. Mooney. (2007). Extracting relations from text from word sequences to dependency paths. In Anne Kao, et al. (Ed.), *Text Mining and Natural Language Processing,* (pp. 29-44). London: Springer.

Culotta, Aron, and Jeffrey Sorensen. (2004). Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Retrieved April 13, 2008, from http://acl.ldc.upenn.edu/P/P04/P04-1054.pdf.

Culotta, Aron, Andrew McCallum, and Jonathan Betz. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics,* (pp. 296-303).

Guy, Marieke, and Emma Tonkin. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine, 12*(1). Retrieved April 13, 2008, from http://www.dlib.org/dlib/january06/guy/01guy.html.

Heymann, Paul, and Hector Garcia-Molina. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems.* Technical Report 2006-10. Department of Computer Science, Stanford University. Retrieved April 13, 2008, from http://labs.rightnow.com/colloquium/papers/tag_hier_mining.pdf.

Iria, Jose, and Fabio Ciravegna. (2005). Relation extraction for mining the semantic web. *Dagstuhl Seminar on Machine Learning for the Semantic Web.* Retrieved April 13, 2008, from http://tyne.shef.ac.uk/t-rex/pdocs/dagstuhl.pdf.

Liu, Hugo and Pattie Maes. (2007). Introduction to the semantics of people & culture (Editorial preface). *International Journal on Semantic Web and Information Systems, Special Issue on Semantics of People and Culture, 3*(1). Retrieved March 28, 2008, from http://larifari.org/writing/IJSWIS2007-SPC-EditorialPreface.pdf.

Mathes, Adam. (2004). *Folksonomies-Cooperative classification and communication through shared metadata.* Unpublished manuscript. Retrieved April 13, 2008, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Mika, Peter. (2005). Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, et al. (Eds.), *Proceedings of the 4th International Semantic Web Conference (ISWC 2005),* (pp. 522–536). Berlin: Springer. Retrieved March 28, 2008, from http://ebi.seu.edu.cn/ISWC2005/papers/3729/37290522.pdf.

Michlmayr, Elke, Sabine Graf, Wolf Siberski, and Wolfgang Nejdl. (2005). A case study on emergent semantics in communities. In Yolanda Gil, et al. (Eds.), *Proceedings of the Workshop on Social Network Analysis, the 4th International Semantic Web Conference (ISWC 2005).* Berlin: Springer.

Nahm, Un Y., and Raymond J. Mooney. (2000). A mutually beneficial integration of data mining and information extraction. *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence,* (pp. 627-632). Menlo Park, CA: AAAI Press.

Nguyen, Dat P., Yutaka Matsuo, and Mitsuru Ishizuka. (2007). Relation extraction from Wikipedia using subtree mining. *Proceedings of the National Conference on Artificial Intelligence Ontology Learning in conjunction with the 14th European Conference on Artificial Intelligence, Berlin, Germany*. Retrieved April 13, 2008, from http://acl.ldc.upenn.edu/N/N07/N07-2032.pdf.

Qin, Jian. (2008). Controlled semantics vs. social semantics: An epistemological analysis. *Proceedings of the 10th International ISKO Conference: Culture and Identity in Knowledge Organization, Montreal, 5.-8. August, 2008.* Retrieved March 28, 2008, from http://web.syr.edu/~jqin/pubs/isko2008_qin.pdf.

Rattenbury, Tye, Nathaniel Good, and Mor Naaman. (2007). Towards automatic extraction of event and place semantics from Flickr tags. In Charles L. Clarke, et al. (Ed.), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* (pp. 103-110). New York: Association for Computing Machinery.

Roth, Dan, and Wen-tau Yih. (2002). Probabilistic reasoning for entity & relation recognition. *Proceedings of 19th International Conference on Computational Linguistics, 1-7.* New Brunswick: ACL.

Sanderson, Mark, and Bruce Croft. (1999). Deriving concept hierarchies from text. In M. Hearst, et al. (Ed.): *Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval,* (pp. 206-213). New York: Association from Computing Machinery.

Schmitz, Patrick. (2006). Inducing Ontology from Flickr Tags. *Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, UK.* Retrieved April 13, 2008, from http://www.topixa.com/www2006/22.pdf.

Shannon, Claude E. (1948). The mathematical theory of communication. *Bell System Technology Journal, 27,* 379-423.

Zelenko, Dmitry, Chinatsu Aone, and Anthony Richardella. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research, 3,* 1083-1106.

The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies

Hak Lae Kim
Digital Enterprise Research Institute, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
haklae.kim@deri.org

Simon Scerri
Digital Enterprise Research Institute, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
simon.scerri@deri.org

John G. Breslin
Digital Enterprise Research Institute, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
john.breslin@deri.org

Stefan Decker
Digital Enterprise Research Institute, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
stefan.decker@deri.org

Hong Gee Kim
Biomedical Knowledge Engineering Lab, Seoul National University
Jong-Ro Gu, Yeon-Gun Dong, Seoul, Korea
hgkim@snu.ac.kr

Abstract

There is a growing interest into how we represent and share tagging data in collaborative tagging systems. Conventional tags, meaning freely created tags that are not associated with a structured ontology, are not naturally suited for collaborative processes, due to linguistic and grammatical variations, as well as human typing errors. Additionally, tags reflect personal views of the world by individual users, and are not normalised for synonymy, morphology or any other mapping. Our view is that the conventional approach provides very limited semantic value for collaboration. Moreover, in cases where there is some semantic value, automatically sharing semantics via computer manipulations is extremely problematic. This paper explores these problems by discussing approaches for collaborative tagging activities at a semantic level, and presenting conceptual models for collaborative tagging activities and folksonomies. We present criteria for the comparison of existing tag ontologies and discuss their strengths and weaknesses in relation to these criteria.

Keywords: tag; tagging; tagging ontology; folksonomy; semantic tagging

1. Introduction

Wikipedia (http://www.wikipedia.com) defines a *Tag* as a 'free-text keyword' and *Tagging* as an 'indexing process for assigning tags to resources'. A *Folksonomy* is described as a shared collection of tags used on a certain platform. The term folksonomy defines a user-generated and distributed classification system, emerging through bottom-up consensus (Vander Wal, 2004). Folksonomies became popular on the Web with social software applications such as social bookmarking, photo sharing and weblogs. A number of social tagging sites such as del.icio.us, Flickr (http://www.flickr.com), YouTube (http://www.youtube.com), CiteULike (http://www.citeulike.org) have become popular.

Commonly cited advantages of folksonomies are their flexibility, rapid adaptability, free-for-all collaborative customisation and their serendipity (Mathes, 2004). People can in general use any term as a tag without exactly understanding the meaning of the terms they choose. The power of folksonomies stands in the aggregation of tagged information that one is interested in. This improves social serendipity by enabling social connections and by providing social search and navigation (Quintarelli, 2005).

The simplicity and ease of use of tagging however, lead to problems with current folksonomy systems (Mathes, 2004). The problems can be classified in two:

- *Local variations*: Tags have little semantics and many variations. Thus, even if a tagging activity can be considered as the user's cognitive process, the resulting set of tags does not always correctly and consistently represent the user's mental model.
- *Distributed variations*: Most tagging systems have their own specific ways of working with and interpreting the meaning of tags. Thus if we want to aggregate tagging data from different applications or services, it's very difficult to find out the meanings and correlations between a sets of tags.

These limitations are due to the lack of a uniform structure and semantic representation found in tagging systems. In this paper, we will compare existing conceptualisations of tagging activities and folksonomies, to assess their merits and thus contribute to future work in this area. Such a conceptualisation, or ontology, is intended to be used in the representation of tagging data in collaborative tagging systems. This paper begins by discussing the reasons why we need Semantic Web technologies for tagging communities. We then briefly overview existing conceptual models for tagging and propose a novel model for folksonomies. We continue by introducing existing tag ontologies and compare them using our conceptual model. Finally, we discuss the results, draw conclusions, and suggest future research areas.

2. Folksonomies: Why Semantic Web Technologies?

2.1. Tagging and Folksonomies

There have been a significant number of efforts to add more structure and semantics to conventional tagging systems. Approaches to tagging and folksonomies have been dominated by a focus on the (statistical) analysis of tag usage patterns (Golder and Herberman, 2006), information retrieval and navigation (Halpin et al., 2006; Jäschke, 2008) and social network analysis and clustering (Mika, 2005; Brooks et al., 2006) based on tagging data. Golder and Herbermann (2005) collected del.icio.us data and analysed the structure and usage patterns of tagging systems. Their work discusses the distinction between collaborative tagging and taxonomies - although collaborative tagging systems have many limitations in terms of semantics and structures, it provides the opportunity to learn from one another through sharing and organising information. Marlow (2006) found that for certain users, the number of tags can become stable over time, while for others, it keeps growing. Cattuto et. al (2007) observed small world effects by analyzing a network structure of folksonomies from Bibsonomy (http://www.bibsonomy.org) and del.icio.us. Their work introduced the notions of clustering and characteristic path length to describe the small world effects. According to the study, folksonomies exhibit a small world structure and have a sort of social network. Mika (2005) carried out a study to construct community-based semantics based on a tripartite model of actors, concepts, and instances. He emphasises the social context for a representation of ontologies and generates the well-known co-occurrence network of ontology learning as well as a novel semantic network based on community relationships using del.icio.us data.

2.2. Semantic Web-Based Approaches

There are a number of debates on the merits of folksonomies when compared to ontologies and other structured vocabulary and classification systems. Despite noted differences between folksonomies and ontologies (Shirky, 2005; Hendler, 2007), Semantic Web technologies can be regarded as a complement to folksonomies. As free-text keywords, tags do not have exact meanings and succumb to linguistic ambiguities and variations including the human error factor. While a user may interpret a tag's semantics through using or reading it, computers cannot automatically understand the meaning, since it is not defined in a machine-readable way (Passant, 2008). Folksonomy systems do not provide a uniform way to share, exchange, and reuse tagging data among users or communities (Kim et al., 2007). With the use of tagging systems in constant increase, these limitations will become evermore critical. As a potential solution, Specia and Motta (2007) propose the integration of folksonomies and ontologies to enrich tag semantics. In particular, Gruber (2007) and Spivack (2005) emphasise the need for folksonomies and ontologies to work together. In general, tag ontologies can contribute in the following three areas:

- *Knowledge Representation Sophistication*: A tag ontology can robustly represent entities and relationships that shape tagging activities. It could make the knowledge structure of tagging data explicit and facilitate the Linked Data (Berners-Lee, 2006) of tagging data on the Web.
- *Facilitation of Knowledge Exchange*: Ontologies enable knowledge exchange among different users and applications by providing reusable constructs. Thus, a tag ontology can be shared and used for separate tagging activities on different platforms.
- *Machine-processable*. Ontologies and Semantic Web technologies in general (knowledge representation, processing and reasoning) expose human knowledge to machines in order to perform automatic data linking and integration of tagging data.

3. Conceptualising Tagging and Folksonomies

Before providing a detailed comparison, we start by reviewing individual conceptual models of tagging activities that preceded our own. A tagging model needs to distinguish between entities in a tagging activity that need to be represented, and address the relationships that exist between them. After reviewing existing tagging models we discuss whether the proposed models are suitable to represent collaborative tagging activities. We then propose our extended model, which caters for the collaborative aspect of folksonomies.

3.1. A Model for Tagging Activities

Many researchers (Mika, 2005; Halpin, 2006; Cattuto, 2007) suggested a tripartite model of tagging activities. Although different authors interpret the term "tagging" differently, we can identify three common entities - users, tags, and resources. They form a triple that represents the Tagging Process:

$$\textbf{Tagging:}\ (U, T, R) \qquad \text{---} \ (1)$$

where U is the set of users who participate in a tagging activity, T is the set of available tags and R is the set of resources being tagged. Gruber (2005) suggested an extension to model (1):

$$\textbf{Tagging}\text{: } (\mathit{object}, \mathit{tag}, \mathit{tagger}, \mathit{source}, + \text{ or } -) \qquad \text{-----------------------} \ (2)$$

where *object*, *tag*, and *tagger* correspond to *R*, *T*, and *U* in the tripartite model. The *source* refers to the tag space where the tagger applies the set of tags whereas the positive/negative parameter is an attempt to represent the collaborative filtering of 'bad' tags from spammers. This tagging model has successfully been used for representing the tagging process at a semantic level. In fact, most tag ontologies have a Tagging class, based on Gruber's model, as a core concept.

3.2. A Model for Collaborative Tagging Activities

Existing models consider tagging as an activity where an individual user assigns a set of tags to a resource. While they provide effective ways to describe the tagging process, they do not really support collaborative tagging activities. We therefore want to provide a *Folksonomy Model* to represent this knowledge, where the folksonomy is considered as a collection of instances of the tagging model. Before doing so, we need to clarify the differences between simple (individual) and folksonomy-based tagging practices. Folksonomies are not created independently by individuals in isolation, but collectively by people who participate in the collaborative tagging activity. Thus, the folksonomy model has to cover all the collaborative aspects and relationships in addition to the objects associated with tagging activities. A straightforward model for a Folksonomy could be defined as follows:

Folksonomy: (*tag set, user group, source, occurrence*) --------------------- (3)

where the tag set is the set of all tags being employed, the user group is a set of users who participate in the tagging activity and the source is the location where the folksonomy is utilised (e.g. social web sites, online communities). The fourth parameter, occurrence, plays an important role to identify the tags' popularity. Comparing this model to the tagging model (2), we can identify the following similarities: the resources (objects) are not part of the Folksonomy model per se. The Folksonomy is rather applied to the collective tagging process of the resources. The tag and tagger parameters in (2) have been replaced with a collective representation of these entities – tag set and user group. The source is still unique since a folksonomy is a multi-user approach to tagging on a single platform. In our opinion, filtering should not be represented at this level. Alternatively, given we represent multiple tags in this model, the frequencies of individual tags become important. Thus, we include the occurrence as our fifth parameter.

Contrary to the concept of Tagging, a folksonomy is a method rather than a process in itself. It can be considered as the practice of acquiring knowledge from collaborative tagging processes. In practice this means that the Folksonomy model should include a representation of the collective tagging processes performed by the group of users. We reflect this in (4) by extending (3) to make the individual tagging activities (to which single users contribute) explicit:

Folksonomy: (*tag set, user group, source, occurrence, Tagging*†) ---------------- (4)

where the last parameter reflects the collective tagging processes performed by the users of the folksonomy, where an individual tagging process is represented by:

Tagging: (*object, tag, tagger*) -- (5)

where object, tag and tagger have the same semantics as those in (2). Thus, our Folksonomy model (4) now incorporates a representation for the collective tagging processes that are individually defined by the Tagging model (5).

4. Overview and Comparison of Tag Ontologies

There is no simple criterion for the comparison of tag ontologies. For this reason, we briefly compare the tag ontologies with respect to their suitability for:

- (a) representing tagging activities and tagging data
- (b) representing features of folksonomies

We will compare seven conceptualisations, keeping in mind the folksonomy model (4) we proposed in Section 3.2. In particular, we include in our comparison a conceptualisation that we presented in our earlier work – the SCOT Ontology (Kim et al., 2008). The choice of the conceptualisations was based on how concrete the model is for tagging and use by online communities. Although a lot of work in analyzing folksonomies has been done in social theory and information retrieval, very few tag ontologies have been reported until today. Few researchers have explicitly specified conceptualisations of tagging data (Borwankar, 2005; Story, 2007) in a formal language. Concerning our selection, at the time of this research only 6 of the 7 conceptualisations were actually proposed as ontologies and described in a dedicated representation language (e.g. OWL). Although Gruber's model is just defined conceptually, we include it in our comparison since many research papers have cited his model and some ontologies have been developed based on this model. The selection of ontologies we include in our comparison (plus Gruber's conceptualisation) is shown in Table 1. Some of the selected conceptualisations better suit the first criterion we have defined at the beginning of this section (a), whereas others are better suited to the second criterion (b). However, all conceptualisations are suitable for both criterions to varying degrees. We will now have a brief look at them individually.

TABLE 7: Features of tag ontologies. **Defined for use in this paper.*

Ontology	URL	Namespace	Format	Update	Applications
Gruber	-	-	-	-	-
Newman	http://www.holygoat.co.uk/projects/tags/	*tags:	OWL	Nov 2005	http://Reyvu.com
Knerr	http://code.google.com/p/tagont/	*tagont:	OWL	Jan 2007	-
Echarte	http://eslomas.com/tagontology-1.owl	*ec:	OWL	2007	-
SCOT	http://scot-project.org	scot:	OWL	June 2008	http://int.ere.st http://relaxseo.com http://openlinksw.com
MOAT	http://moat-project.org	moat:	OWL	Feb 2008	http://openlinksw.com lord.info
NAO	http://www.semanticdesktop.org/ontologies/nao/	nao:	NRL	Aug 2007	Nepomuk

Gruber's work is an early attempt to conceptualise tagging activities. His model can be viewed as a first step towards a general applicable representation model for tagging. Although his model itself is not an ontology it clearly reveals a generic conceptualisation of tagging. For more details on his work we refer to Gruber (2007, 2008). Newman's model (referred to as Newman) describes relationships between an agent, an arbitrary resource, and one or more tags. In this model there are three core concepts such as Tagger, Tagging, and Tag to represent a tagging activity. Knerr (2006) provides the tagging concept in the Tagging Ontology (referred to as Knerr) and Echarte et. al (2007) propose a model for folksonomies (referred to as Echarte). Since their approaches are based on the ideas of Gruber and Newman, the core elements of the ontologies are almost identical. In particular, Echarte's model extends concepts such as time, domain, visibility, type, etc., and is represented by OWL. The SCOT Ontology - Social semantic Cloud of Tags, describes the structure and semantics of tagging data and enables interoperability of tagging data among heterogeneous social websites and tagging applications. Although SCOT's main goal is to represent collaborative tagging activities, it is also suitable for representing the features of folksonomies (e.g. source, user group, frequencies, tag co-occurrence, etc.). MOAT (Passant, 2008) - Meaning of a Tag, is intended for semantic-annotation of content by providing a meaning for free-text tagging. In addition to extensions to the Tag, Tagging, and Tagger concepts from Newman's ontology, MOAT provides the Meaning class to represent custom, user-provided 'meanings' for tags. The Nepomuk Annotation Ontology (NAO) (Scerri et. al, 2007) is provided for annotating resources on the Social Semantic Desktop (http://www/nepomuk.semanticdesktop.org/). It is not entirely dedicated to tagging practices but demonstrates the increasing importance of tagging representation in social systems.

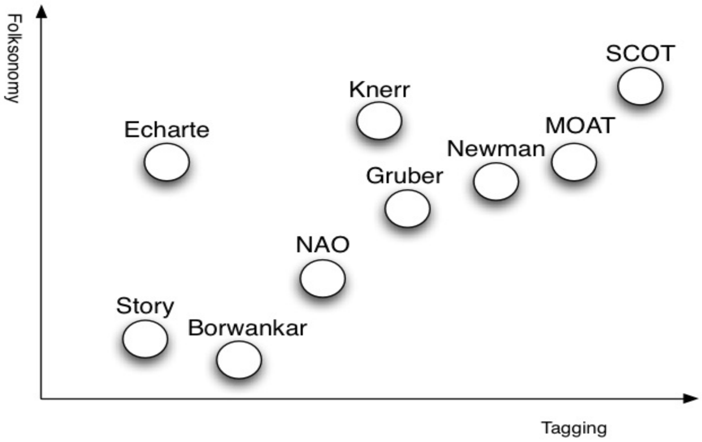


FIG. 4: Criterion suitability for different conceptualisations.

FIG. 4 demonstrates the different inclinations of the seven conceptualisations listed in Table 1, given the criteria discussed at the start of this section. Whereas Newman’s Ontology is more inclined towards representing tagging data, and Echarte’s Ontology towards representing features of folksnomies, SCOT has a higher level of sophistication in both directions. In the next section, we will detail the main entities and features that the six ontologies and Gruber’s model are able to represent. We will support our conclusions in this section by exploring the suitability of the individual conceptualisations, vis-à-vis criteria (a) and (b) as set out in the start of this section. We start by listing and comparing the concepts (classes) and proceed by listing and comparing their features (attributes).

4.1. Class Comparison

In this section we discuss in more details the general comparison we presented in the previous section. First, we will have a look whether the individual conceptualisations are suitable for representing general tagging activities and tagging data. All models have a representation for the o*bject*, *tag* and *tagger* in our Tagging model (5) and all except NAO have a concept representing the tagging process. In Newman’s model, the tagging concept is further refined into *tags:RestrictedTagging* (exactly one tag for a resource) and *tags:Tagging* (one or more tags for a resource). Echarte et al. provide the *Annotation* class to represent the tagging activity – i.e., it is the same as *tags:Tagging*. Thus, the *Tagging* concept can be considered as a core concept of tag ontologies. Although SCOT and MOAT have different goals compared to others, they also can describe tagging by linking to the tags:Tagging class in Newman’s ontology.

TABLE 8: Ontology concepts. Concepts are locally defined unless otherwise stated (e.g. rdfs:Resource).

Model	Resource	Tag	Tagging	Tag Set	User	User Group	Source	Others
Gruber	*Object*	*Tag*	*Tagging*		*Tagger*		*Source*	*Polarity*
Newman	rdfs:Resource	:Tag	:Tagging		foaf:Agent			:RestrictedTagging
Knerr	rdfs:Resource	:Tag	:Tagging		:Tagger	foaf:Group	:Service Domain	:VisibilityEnum
Echarte	:Resource	:Tag	:Annotation		:User		:Source	:Polarity
SCOT	sioc:Item	:Tag	tags:Tagging	:TagCloud	sioc:User	sioc:Usergroup	sioc:Site	:Cooccurrence
MOAT	rdfs:Resource	tags:Tag	tags:Tagging		foaf:Agent			:Meaning
NAO	rdfs:Resource	:Tag			:Party			

We now consider whether the ontologies address collective tagging data and provide sufficient features of folksonomies, as described in our Folksonomy model (5). Some ontologies which are based on Gruber’s model (which was not designed for folksonomies) have been extended in order to support folksonomies. For instance, Knerr and Echarte introduce the *ServiceDomain* and the

Source class to represent the *source*. In addition, Knerr allows a user to use *foaf:Group* alongside *foaf:Person* to describe the *user group*. Similarly, NAO allows the user to use nao:Party to represent the *user group*. MOAT does not have a class for defining it. Nevertheless they are not enough to represent folksonomies at a semantic level. SCOT is consistent with the folksonomy model and provides representations for the *source*, *user group* and *tag set*. In Table 2 we compare the classes provided by these conceptualisations that are relevant to our study. Additionally, we must note that although an ontology might not provide all the required representations, they can act as a "good Semantic Web citizen" by connecting to external vocabularies such as SIOC (Semantically-Interlinked Online Community), FOAF (Friend-of-a-Friend), SKOS (Simple Knowledge Organisation System), or DC (Dublin Core Metadata) to further weave data on the Web. For example, MOAT and SCOT use the SIOC ontology extensively to describe online communities, while other ontologies do not reuse or link to external terms. In particular, although Echarte has its own classes to represent a tagging and a folksonomy, the classes do not have any relations with other RDF vocabularies.

TABLE 3: Data type properties. The table shows value attributes for some core concepts, interpreted as domain (row) – property – range (column).

	Literal	Time	Numeric Values
Source	tagont:hasServiceName		
Resource	ec:hasURI ec:hasSourceName		
User	ec:hasUserName		
Tag Set	dc:title dc:description	scot:updated	scot:totalTags scot:totalTagFrequency scot:totalItems scot:totalCooccurTags scot:totalCooccurFrequency
Tag	tags:name tags:tagName tagont:prefTagLabel tagont:hasTagLabel nao:prefSymbol nao:prefLabel nao:description ec:hasPrefLabel ec:hasLabel ec:hasAltLabel ec:hasHiddenLabel	scot:lastUsed nao:created nao:lastModified	scot:ownAFrequency scot:ownRFrequency scot:cooccurAFrequency scot:cooccurRFrequency ec:hasPosition
Tagging	tagont:hasNote	tags:taggedOn tagont:isTaggedOn ec:hasDateTime	

4.2. Attribute Comparison

While the number of classes enhances taxonomical representations, the power of ontologies lies in the ability of representing relationships between the classes. Although most of the studied ontologies have a similar taxonomical structure, their attributes vary according to their goals and purposes. We will now have a look at the attributes provided by the ontologies, and compare their functionalities. We differ between data type attributes, which relate classes to non-conceptual data (e.g., string or date), and object type properties which provide relationships between classes.

Data Type. Aside from declarative features that represent relationships among users, tags, and resources, a semantic model for folksonomies needs to provide for descriptive features that state non-conceptual values. Most surveyed tag ontologies have many attributes to describe data-type values, i.e. numerical quantities, free-text descriptions, date, time, etc. The data-type properties relevant to this work are summarised in Table 9. A number of datatype properties are either directly or indirectly (i.e. via subPropertyOf) reused from the Dublin Core vocabulary. For instance Newman's ontology *tags:name* is a subproperty of *dc:title* and *tags:taggedOn* is a subproperty of *dc:date*. Only SCOT provides for the description of numerical values for entities, e.g. *scot:totalTags* (attributed to a *scot:TagCloud*) refers to the total number of tags in a tag cloud and *scot:totalItems* refers to the total number of resources tagged with tags in the tag cloud. SCOT also provides properties relating to the frequency of a tag itself. Whereas the simplistic *scot:ownAFrequency* refers to the actual occurrence(s) of a particular tag in a tag cloud, *scot:ownRFrequency* represents the percentage frequency of a tag within a particular tag cloud, relative to the total of all tag frequencies in that tag cloud.

There are many attributes to describe string and literal values for a specific purpose, e.g. *tags:name*, *tagont:prefTagLabel*, *nao:preLabel*, and *ec:hasLabel* for describing tag's name.

Table 4: Object type properties. The table shows relationships between core concepts, interpreted as domain (row) – property – range (column).

	Source	*Resource*	*User Group*	*User*	*Tag Set*	*Tag*	*Tagging*	*Others*
Source								tagont: hasServiceHomepage ec:hasSource
Resource						tags:taggedWithTag scot:hasTag nao:hasTag	tags:tag	
User Group							tagont:hasTagging	
User							tagont:hasTagging	
Tag Set	scot:tagSpace		scot:hasUsergroup	scot:createdBy	scot:composedOf	scot:contains	scot:taggingActivity	
Tag		tags: isTagOf scot:tagOf nao:isTagFor ec:hasRelatedResource		scot:usedBy nao:creator	scot:containedIn	tags:equivalentTag tags:relatedTag scot:aggregatedTag scot:spellingVariant scot:delimited tagont:sameTag ec:hasTag		moat:hasMeaning ec:hasPolarity scot:cooccursIn scot:cooccursWith
Tagging	tagont: hasServiceDomain ec:hasSource	tags:taggedResource tagont:hasTaggedResource ec:hasResource moat:tagMeaning		tags: taggedBy tagont:hasTagger ec:hasUser		tags:associatedTag tagont:hasTag ec:hasAnnotationTag		tagont:hasType tagont:hasVisibility

Object Type. The object type properties relevant in the context of this study are summarised in Table 4. SCOT, Echarte and Knerr provide the possibility to define a tagging activity. In SCOT, there is no local property to describe who is involved in a tagging activity. For this purpose SCOT reuses Newman's *tags:taggedBy* attribute. Via SCOT one can describe who uses tags via the *scot:usedBy* property. Meanwhile, three ontologies have the property to identify a location or source in which the tagging occurred. TagOnt provides *tagont:hasServiceDomain* to link the tagging activity to the *ServiceDomain*, Echarte provides *ec:hasSource* with the *Source* as its range value, whereas SCOT provides *scot:tagspace* with a range of *sioc:Site*. The relation between tags and resources is defined via tags:isTagOf (range: rdfs:Resource), nao:isTagFor (range: rdfs:Resource), and scot:tagOf (range: sioc:Item) properties in theNewman, NAO and SCOT ontologies respectively. They also provide inverse properties for this relation. Defining relations between tags is one of the benefits of using an ontology to model folksonomies, since this effectively gives semantics to tags in a tag set. Nevertheless only SCOT and Newman take advantage of this possibility. Whereas Neman provides very restricted properties such as *tags:equivalentTag* and *tags:relatedTag*, SCOT provides many more attributes such as *scot:spellingVariant* and *scot:delimited*. The spelling variant property is further refined into *scot:acronym*, *scot:plural*, *scot:singular* and *scot:synonym*. In addition, the latter has further subproperties to define specific synonym types, i.e. *scot:hypenatated*, *scot:underscored*, *scot:slashed*, and *scot:spaced*. In comparison to other ontologies, SCOT specifically provides attributes that represent characteristics of folksonomies such as *scot:hasUsergroup*, *scot:createdBy*, *scot:contains*, and *scot:taggingActivity*.

To conclude this section we briefly give a summary of the comparison. So far, tag ontologies have mainly been used for representing tagging activities, and only to a minor extent for modeling the features of folksonomies. According to the Folksonomy model given in Section 3.2, SCOT is suitable for this model. But, we might argue that the surveyed ontologies have different ontological purposes and different expressivity. Therefore, as an ideal solution we might need to interlink among the proposed ontologies.

5. Conclusion

In the first half of this paper we proposed a model for collaborative tagging activities and folksonomies – based on the widely accepted model for tagging. The detailed comparisons presented in Section 4 support several general concluding observations about ontologies related to tagging activities and their usefulness in collaborative tagging systems. This research can be considered as a first attempt to systematically compare different conceptualisations of semantic tagging for collaborative tagging systems. We believe that tag ontologies should be evaluated with respect to a particular goal, application or scenario rather than merely for the sake of an evaluation. Our observations take into consideration two separate criteria – the depth of tagging data per se, and the collaborative aspect in folksonomies. As we mentioned in the start of the paper, tag ontologies are in an early stage and current approaches need to be elaborated or combined to enrich schemas and meet both criteria. Nevertheless the surveyed ontologies already offer an improved opportunity for collaborative tagging systems – especially given the machine-processable representations that they can provide.

Following the comparison of the tag ontologies we arrived at the following conclusions:

- There is agreement on the issue as to what are the most elementary building blocks of a model for the tagging. The building blocks consist of the taggers, the tags themselves, and the resources being tagged.
- Different individuals create substantially different conceptualisations of tagging data and tagging activities despite the fact that their purposes are similar.
- The tag model does not cover overall characteristics of a folksonomy. SCOT, combined Gruber's conceptual model and Newman's vocabularies, is the ontology that must be suitable to represent collaborative tagging activities and it provides the most appropriate representations for the Folksonomy model as we defined it. In addition linking between SCOT and MOAT is useful way to complement to define a meaning of tag.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

References

Berners-Lee, Tim. (2006). *Linked Data.* Retrieved June 14, 2008, from http://www.w3.org/DesignIssues/LinkedData.html.

Borwankar, Nitin. (2005). *TagSchema: Slicing and dicing data 2.0: Foundation Data Model for Folksonomy Navigation.* Retrieved June 14, 2008, from http://tagschema.com/blogs/tagschema/2005/06/slicing-and-dicing-data-20-part-2.html.

Brooks, Cristopher. H., and Nancy Montanez. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. in WWW 06. *Proceedings of the 15th international conference on World Wide Web,* (pp.625-632.). New York: ACM Press.

Cattuto, Ciro, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, and Andreas Hotho, et. al. (2007). Network Properties of Folksonomies. *AI Communications 20*(4), 245 - 262.

Echarte, Francisco, José J. Astrain, Alberto Córdoba, and Jesús Villadangos. (2007). Ontology of folksonomy: A New modeling method. *Proceedings of Semantic Authoring, Annotation and Knowledge Markup (SAAKM).*

Golder, Scott A., and Bernardo A. Huberman. (2006). The structure of collaborative tagging systems. *Journal of Information Sciences, 32*(2), 198-208.

Gruber, Tom. (2007), Ontology of folksonomy: A mash-up of apples and oranges. *Int. Journal on Semantic Web and Information Systems, 3*(2).

Gruber, Tom. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics 6*(1), 4-13.

Halpin, Harry., Valentin Robu, and Hana Shepard. (2006). The dynamics and semantics of collaborative tagging. *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW06).*

Hendler, James. (2007, November 11). Shirkyng my responsibility. Message posted to http://www.mindswap.org/blog/2007/11/21/shirkyng-my-responsibility/.

Jäschke, Robert, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme. (2008). Discovering shared conceptualisations in folksonomies. *Web Semantic, 6*(1), 38-53.

Kim, Hak-Lae., Sung-Kwon Yang, John G. Breslin, and Hong-Gee Kim. (2007). Simple algorithms for representing tag frequencies in the SCOT exporter. *IAT Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, (pp. 536-539).

Kim, Hak-Lae, Alexandre Passant, John G. Breslin, Simon Scerri, and Stefan Decker. (2008). Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. *Proceedings of the 2nd International Conference on Semantic Computing, San Francisco, USA.*

Knerr, Torben. (2006). *Tagging ontology- towards a common ontology for folksonomies*. Retrieved June 14, 2008, from http://tagont.googlecode.com/files/TagOntPaper.pdf.

Marlow, Cameron, Moor Naaman, Danah Boyd, and Mark Davis. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. *Proceedings of the seventeenth Conference on Hypertext and Hypermedia,* (pp. 31-40). New York: ACM Press.

Mathes, Adam. (2004). *Folksonomies - Cooperative classification and communication through shared metadata.* Retrieved June 14, 2008, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Mika, Peter. (2005). Ontologies Are Us: A unified model of social networks and semantics. *Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland,* (pp. 522-536). Berlin/ Heidelberg: Springer.

Miles, Alistair, and Dan Brickley. (2005). *SKOS core vocabulary specification*. Retrieved June 14, 2008, from http://www.w3.org/TR/swbp-skos-core-spec.

Newman, Richard, Danny Ayers, and Seth Russell. (2005). *Tag ontology*. Retrieved June 14, 2008, from http://www.holygoat.co.uk/owl/redwood/0.1/tags/.

Passant, A. (2007). Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. *Proceedings of International Conference on Weblogs and Social Media.*

Passant, A., and Laublet. P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and Linked Data. *Proceedings of Linked Data on the Web (LDOW 2008).*

Quintarelli, Emanuele. (2005). *Folksonomies: Power to the people.* Retrieved June 14, 2008, from http://www.iskoi.org/doc/folksonomies.htm.

Scerri. Simon, Michael Sintek, Ludger van Elst, and Siegfried Handschuh. (2007). *NEPOMUK Annotation Ontology.* Retrieved June 14, 2008, from http://www.semanticdesktop.org/ontologies/nao/.

Shirky, Clay. (2005). *Ontology is overrated: Categories, links, and tags.* Retrieved June 14, 2008, from http://www.shirky.com/writings/ontology_overrated.html.

Specia, Lucia, and Enrico Motta. (2007). Integrating folksonomies with the semantic web. *European Semantic Web Conference, 2007, Innsbruck, Austria,* (pp. 624-639). Retrieved from http://www.eswc2007.org/pdf/eswc07-specia.pdf.

Spivack, Nova. (2005, January 21). Folktologies -Beyond the folksonomy vs. ontology distinction. Message posted to http://novaspivack.typepad.com/nova_spivacks_weblog/2005/01/whats_after_fol.html.

Story, H. (2007, February 6). Search, tagging and wikis. Message posted to http://blogs.sun.com/bblfish/entry/search_tagging_and_wikis.

Wal, Thomas V. (2004). *Folksonomy.* Retrieved June 14, 2008, from http://vanderwal.net/folksonomy.html.

Project Reports

Session 1:

Toward the Semantic Web

DCMF: DC & Microformats, a Good Marriage

Eva Méndez
University Carlos III of Madrid, Spain
emendez@bib.uc3m.es

Leandro M. López
Freelance, Argentina
inkel.ar@gmail.com

Arnau Siches
esbudellat.net, Spain
asiches@gmail.com

Alejandro G. Bravo
Webposible, Spain
alejandrogbravo@yahoo.es

Abstract

This report introduces the Dublin Core Microformats (DCMF) project, a new way to use the DC element set within X/HTML. The DC microformats encode explicit semantic expressions in an X/HTML webpage, by using a specific list of terms for values of the attributes "rev" and "rel" for <a> and <link> elements, and "class" and "id" of other elements. Microformats can be easily processed by user agents and software, enabling a high level of interoperability. These characteristics are crucial for the growing number of social applications allowing users to participate in the Web 2.0 environment as information creators and consumers. This report reviews the origins of microformats; illustrates the coding of DC microformats using the Dublin Core Metadata Gen tool, and a Firefox extension for extraction and visualization; and discusses the benefits of creating Web services utilizing DC microformats.

Keywords: microformats; Dublin Core; DCMES; Web 2.0; metadata; X/HTML; RDF; embedded Web semantics; social applications; bibliographic data repositories

1. Introduction

During the Web 1.0 years ("Altavista Age" that you probably remember), the usual method for including semantic information within documents was using the (X)HTML header <meta> elements, as well as <title>, <address>, <link>, <del>, <ins> elements and "title" and "cite" attributes. This continues in the present, but the abuse ("black SEO") and misuse (inconsistencies) of <meta> elements forces search engines to ignore this information. With the introduction and growing popularity of XML, and the first Recommendation status of W3C's RDF in February 1999 (W3C, 1999b), the potential and versatility of metadata has increased tremendously, supporting more precise and interoperable information gathering and retrieval. The Semantic Web aims to transform the current Web into a machine-readable Web, while maintaining its ability to be directly and easily read by people. However, metadata in webpages is not person-oriented, but search engine-oriented. This metainformation is only available through visualizing source code or using metadata visualization tools (Firefox Dublin Core Viewer extension (2005), or, historically, using special user agents like Metabrowser, which allowed the user to browse both the information and the metainformation within a webpage.

2. Microformats

Microformats originated from a grassroots movement lead by Tantek Çelik to make recognizable data items (such as events, contact details or geographical locations) capable of automated processing by software agents, as well as directly readable by human beings (Knowledge@Wharton, 2005). The official website of microformats.org says that they are *"designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards".*

A microformat is a Web-based data formatting approach seeking to re-use existing content as metadata, using only X/HTML classes and other attributes for encoding. Microformats are simple conventions for embedding semantic markup in human-readable documents. They make use of implicit and explicit X/HTML characteristics in order to add simple semantic information via:

- relationship links using "rel" and "rev" attributes on <a> and <link> elements. Besides the default defined types of relationship in the HTML specification, they can also be extended using profiles.
- "class" and "id" attributes of most X/HTML elements. In this case, in addition to its support for display (as in CSS), these attributes may be used for other different functionalities.

Web developers frequently make use of meaningless values for class names and identifiers. However, source code comprehension can be enhanced and *extra information* added for instance using "header", "menu" and "footer" for page layout definition. In December 2005 Google did an analysis of a sample of slightly over a billion documents, extracting information about popular class names, elements, attributes, and related metadata. One of the goals of that project (Google Web Authoring Statistics) was to know if any logic or semantics were used in class names. The conclusion was that there is no uniformity in naming classes. As a consequence, it is hard to parse documents in order to extract semantic information, except when microformats are used.

The main goal of microformats is to solve problems created by inconsistent labeling, for instance, defining events, people, relationships, etc. through the creation of simple elements and element sets. Some of the microformat element sets are associated with widely adopted standards or schema, such as hCard (based on the vCard standard for business cards) and hCalendar (based on iCalendar for events); some others have a newer origin, like "rel-tag" microformats, used to simplify blog indexing through Technorati. There are also other globally used microformats such as "vote-links" for electronic voting, "hReview" for media reviews, "hResume" for resumes, and "XFN" for social networks, etc.

One of the most obvious and important benefits of using microformats —besides easy encoding and quick distribution— is the ability to easily parse web documents to look for microformats and extract them. There are a number of Web services that exploit this semantic information such as: Technorati[28] to find Weblog posts, Upcoming.org[29] to extract hCalendar definitions of events, and Yahoo! Tech[30] publishing of products reviews etc. Yahoo! has also implemented a search engine for Creative Commons licensed documents[31], and Yahoo! Search parses almost every defined microformat.

3. Dublin Core Microformats (DCMF)

We started the Dublin Core MicroFormats (DCMF) project in 2005, taking advantage of Dublin Core's versatility, general purpose applicability, its formal standardization and the wide promotion by the Dublin Core Metadata Initiative (DCMI). Dublin Core is a metadata schema which is syntactic-independent so it is suitable for encoding semantics within a microformats structure. So, DCMF allow us to extend the indisputable advantages of DCMI —simplicity, flexibility, diffusion and appropriateness— to any domain. All of the microformats have been created with a concrete goal, and the general goal of DC Microformats is to describe web resources (as any resources can have a title, keywords, description, author, etc.). But DC microformats are also particularly appropriate to encode bibliographic descriptions of resources, such as magazines, books, articles, in any media, including paper or digital.

[28] Technorati: http://technorati.com/

[29] Upcoming: http://upcoming.yahoo.com

[30] Yahoo! Tech: http://tech.yahoo.com

31 Yahoo Search: Creative Commons Search: http://search.yahoo.com/cc

3.1. Example: DCMF Encoding

Let's see an example of how we can describe Tim Berners-Lee's book using semantic information encoded as DCMF. The following code will represent this information in an X/HTML webpage:

```
<dl class="dublincore">
<dt>Title:</dt>
<dd class="title">Weaving the Web</dd>
<dt>ISBN:</dt>
<dd class="identifier">0062515861</dd>
<dt>Author:</dt>
<dd><a href=”http://www.w3.org/People/Berners-Lee” class="creator">Tim
Berners-Lee</dd></dl>
```

According to the example, to use DC microformats, we need:

1. An X/HTML element (in this example <dl>, a definition list) with the class or identifier "dublincore", which acts as container of a DC microformat and identifies it.
2. A string which represents the semantic expressed by the microformat (in the example, “Title”, “ISBN” and “Author”).
3. An X/HTML element with the "class" or "id" attributes, whose value is the appropriate DC element to indicate the semantic information to machines (in the example “title”, “identifier” and “creator”); and also the value of the element/property (in the example “Weaving the Web”, “0062515861” and “Tim Berners-Lee”).

If we declare the information expressed in the microformat (Web 2.0 approach) in RDF nomenclature (Semantic Web approach), we should speak about: resource, property and value, where:

- resource, is the value of the element with "identifier" class or id, if it exists;
- property, is the value of the class expressed for both; for humans (“Title”, “ISBN” and “Author”) and for machines (“title”, “identifier” and “creator”)
- and value, is the content of X/HTML elements with the class or identifiers of the last item (“Weaving the Web”, “0062515861” and “Tim Berners-Lee”).

3.2. How to Create DCMF: Dublin Core Metadata Gen

There are many tools to extract and/or generate metadata with DCMI elements, but none allows us to create microformats, except Dublin Core Metadata Gen, which was incorporated into the DCMF project. Dublin Core Metadata Gen is an application developed in PHP that generates three kinds of DC metadata: RDF, X/HTML using <meta> elements and also, per the project presented here, DCMF. In Dublin Core Metadata Gen, you can enter the data into a template and get: DC in RDF, DC in X/HTML using <meta> elements, and DC in microformats.

3.3. How to See DCMF: Dublin Core Microformats Viewer

The Dublin Core Microformats Viewer is an add-on for Firefox and Flock browsers. This user agent's extension detects DC microformats when is then included in the X/HTML code of the webpage. Like Dublin Core Viewer Extension add-on (and inspired on it), DC Microformats Viewer installs a little icon in the status bar, letting the users open a pop-up window containing a table with the Dublin Core microformats present in the current page. This tool is only a simple extension with simple functionality, but it also shows the ease of extracting metainformation from DC microformats, and the potential of this approach.

4. Microformats, <meta> Elements and/or RDF

Microformats are another way of expressing metadata in general, and DC in particular, embedded in web resources. If we compare microformats encoding with the use of <meta>

elements, and/or with RDF syntax, microformats have some advantages and distinctive characteristics:

- Easy to create. Microformats make participation in Web 2.0 social collaboration easier for content creators. Any web content creator can write microformats easily. The only required knowledge is basic X/HTML and X/HTML authoring tools.
- Easy to recognize and use. The information (of an event, a business card, bibliographic record, etc.) can be read by people using their user agents. Users also can extend their browsers' functionalities (mainly by add-ons and widgets, such as Operator for Firefox), to combine pieces of information on websites with applications (e.g. Flickr+Google Maps; Upcoming+ Google Calendar; Yahoo! Local+your address book, etc.).

There are also disadvantages. Probably microformats are less known than the <meta> element, because microformats belong to the emerging domain of the Web 2.0. Also microformats are more limited than RDF; for example, they can not formally define complex relationships and microformats' scope are narrower that the descriptive potential of RDF. Despite all those limitations, microformats are a way to work with DC metadata in the context of Web 2.0, allowing authors to generate semantic information easily comprehensible to both people and machines. Web services can also be developed to support DC microformats, as for any other existing microformats. Examples might include article repositories, books, magazines, etc that allow people to add and find bibliographic records easily.

Microformats, have been also called as "lower-case Semantic Web" but they are a very important inflection point within the Semantic Web. Standards like GRRDL, a recent W3C Recommendation, demonstrate that mechanisms from *Gleaning Resource Descriptions from Dialects of Languages,* are needed to extract Semantic Web Information from X/HTML microformats (W3C, 2007).

5. Conclusions and Future Work

In a post on his blog, Stu Weibel (2006) wrote: *The flexibility that microformats afford is an essential feature of the hyper-innovation that characterizes Web 2.0,* but he wondered if Dublin Core fits in the microformats' philosophy. In this report we answer "YES": Dublin Core fits perfectly in the microformats' philosophy, just as it does in the context of the "classic" Semantic Web. Adopting microformats as a new way to express semantic information with DC allows us to expand the use of DC to new domains that, otherwise, would not use it. In addition, the nature of DC as a general purpose metadata model implicitly suggests its use in microformats for describing resources, especially the bibliographic types of resources previously mentioned.

Microformats avoid the problems of updating and synchronizing the information in many sources (like resumes on employment-related websites) or formats (information visible for people in Web pages, or <meta> elements and RDF for search engines). But microformats especially are intended to allow people to participate in and take advantage of the Semantic Web in the specific situations already mentioned.

The DCMF project intends to combine the simplicity and flexibility of Dublin Core with the possibilities that microformats offer. DCMF is an attempt to make semantic information easy and practical. Furthermore, the ease of parsing web documents with microformats lets us use this semantic information for Web services useful to people.

Future work on DC microformats will be the evolution and improvement of those tools described here (Dublin Core Metadata Gen and Viewer and Dublin Core microformats), and the development of Web services for querying the information within DC microformats.

References

Bravo, Alejandro, and Arnau Siches. (2005). *Dublin Core Metadata Gen: Generator of metadata using Dublin Core.* Retrieved from http://www.webposible.com/utilidades/dublincore-metadata-gen.

Conolly, Dan. (2007). *Gleaning Resource Descriptions from Dialects of Languages (GRDDL) W3C Recommendation 11 September 2007.* Retrieved, April 1, 2008, from http://www.w3.org/TR/2007/REC-grddl-20070911/.

DCMF. (2008). *Microformatos Dublin Core* (Translated to Spanish). Retrieved from http://webposible.com/microformatos-dublincore/.

DCMI. (2008). *Dublin Core Metadata Initiative.* Retrieved from http://dublincore.org.

DCMI. (2008). *Dublin Core Metadata Initiative - Tools and Software*. Retrieved from http://dublincore.org/tools.

Google. (2006). *Web Authoring Statistics.* Retrieved from http://code.google.com/webstats/.

Kaply, Michael. (2008, May 21). Operator 0.9.3 for Firefox. Posted to https://addons.mozilla.org/es-ES/firefox/addon/4106.

Knowledge@Wharton. (2005, July 27). *What's the Next Big Thing on the Web? It May Be a Small, Simple Thing - Microformats.* Retrieved, April 1, 2008, from http://knowledge.wharton.upenn.edu/index.cfm?fa=printArticle&ID=1247.

Kumar, Amit. (2008, March 13). The Yahoo! Search Open Ecosystem. Message posted to http://www.ysearchblog.com/archives/000527.html.

Lauke, Patrick H. (2005). *Firefox Dublin Core Viewer Extension.* Retrieved from http://www.splintered.co.uk/experiments/73/.

López, Leandro M. (2008, March 19). Visor de Microformatos Dublin Core: Extensión para Firefox. Message posted to http://wses.wordpress.com/2008/03/19/visor-de-microformatos-dublin-core.

Metabrowser. (n.d.). Retrieved from http://metabrowser.spirit.net.au. (Metabrowser is not longer available).

Ora, Lassila, and Ralph R. Swick. (1999b). *Resource Description Framework (RDF) Model and Syntax Specification W3C Recommendation 22 February 1999*. Retrieved, April 1 2008, from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Raggett, Dave, Arnaud Le Hors, and Ian Jacobs (Eds.). (1999a). *HTML 4.01 Specification W3C Recommendation 24 December 1999.* Retrieved, April 1 2008, from: http://www.w3.org/TR/html401.

Weibel, Stuart. (2006, April 12). Ockham's Bathroom Scale, Lego™ blocks, and Microformats. Message posted to http://weibel-lines.typepad.com/weibelines/2006/04/ockhams_bathroo.html.

Making a Library Catalogue Part of the Semantic Web

Martin Malmsten
National Library of Sweden,
LIBRIS department, Sweden
martin.malmsten@kb.se

Abstract

Library catalogues contain an enormous amount of structured, high-quality data, however, this data is generally not made available to semantic web applications. In this paper we describe the tools and techniques used to make the Swedish Union Catalogue (LIBRIS) part of the Semantic Web and Linked Data. The focus is on links to and between resources and the mechanisms used to make data available, rather than perfect description of the individual resources. We also present a method of creating links between records of the same work.

Keywords: rdf; library catalogue; semantic web; linked data; persistent identifiers; frbr; sparql

1. Introduction

Even though bibliographic exchange has been a reality for decades, exchange of authority information and links between records are still not widely implemented. The standard way of making bibliographic data available is still through search-retrieve protocols such as SRU/W[32] or Z39.50[33]. Though this makes single bibliographic records retrievable, it does not provide a way to directly address them and reveals little or nothing about links between records. In contrast the Semantic Web (Berners-Lee et al., 2001) is by definition built upon linking of information. The promise of the Semantic Web and Linked Data (Berners-Lee 2006) is that it could make data connected, simply by making it available. This, it seems, could be the perfect way for libraries to expose all of their data.

A goal when creating the new version of the LIBRIS web interface[34] was to make the information presented to a normal user transparently available to machines/web robots as well. It was also obvious that information not intrinsic to the record itself, such as user annotations and connections to other records could be made available this way.

Also, thirty years of continually changing cataloguing rules and practices have left some data in an inconsistent state. Our hope is that the result of the work described will help us work with data in a new and better way.

2. Technical Overview

The Swedish Union Catalogue comprises about 175 libraries using a single Integrated Library System (ILS) for cataloguing. MARC21 is used for bibliographic, holdings and authority records. It contains about six million bibliographic records. A number of components were developed to make the ILS “talk RDF”.

We created an RDF server wrapper to make the ILS accessible through HTTP and able to deliver RDF describing bibliographic and authority resources upon request, as well as RDF describing the links between them. Persistent URIs were created by using each record’s unique number, these URIs can be dereferenced and will deliver the RDF when queried properly through HTTP content negotiation.

[32] Search/Retrieval via URL - http://www.loc.gov/standards/sru/
[33] Z39.50 - http://www.loc.gov/z3950/agency/
[34] LIBRIS - http://libris.kb.se/

This data could then be loaded into a triple store to enable searching using SPARQL (Prud’hommeux and Seaborn, 2008).

3. Implementation

In this section we will outline the individual components of the implementation. A schematic is provided in FIG. 1.

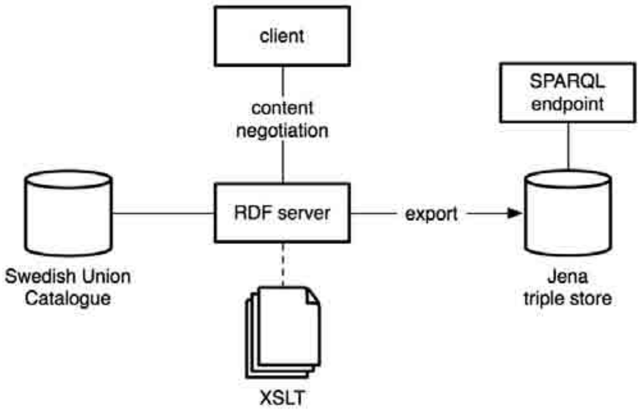


FIG. 1. Implementation schematic

3.1. RDF Server Wrapper

The first step was to create a wrapper around the ILS that could deliver the records in RDF rather than the binary format normally used for bibliographic records (ISO2709). The wrapper talks to the ILS using SQL and delivers records given its unique number. It then converts the ISO2709 record into an XML representation of MARC21. In the final step a transformation is applied to the XML using XSLT (Clark 1999).

Since each output format is implemented in a single XSLT-file, adding a new format or making changes to an existing one is trivial.

3.2. Linked Data and Access

Links and access are crucial underpinnings of both the semantic and “normal” web. For a resource to be linkable it needs a URI, for it to be accessible, that URI should be a HTTP one. Following the four rules of Linked Data (Berners-Lee 2006), a persistent, dereferenceable URI is created for each record. For bibliographic records: http://libris.kb.se/resource/bib/<number>, and for authority records: http://libris.kb.se/resource/auth/<number>.

Using HTTP content negotiation, the correct format can be delivered depending on the clients capabilities. This method uses the HTTP Accept header to tell the server what media types the client can handle and prefers. For example, the accept header text/html tells the server to deliver an HTML page suitable for a human user. An accept header containing, for example, text/rdf+n3 or application/rdf+xml tells the server that the client is able to handle RDF. The server can either deliver the data in RDF directly or send an HTTP 302 or 303 response indicating that the information can be found at a different URL (Sauermann, Cyganiak, 2008). See Appendix A for an example of content negotiation.

3.3. SPARQL Endpoint

We were interested in using SPARQL as a tool to both query and analyze data. Some queries that can be hard, or impossible, to formulate using SQL or a full text search language are easily formed using SPARQL. For example, the following query: “show me all subjects of records that belongs to the same work as the record with identifier XYZ". A query like this can be very useful for someone wanting to "auto complete" missing subject entries on records belonging to the same work. We used the Jena Semantic Web Framework[35] to create a triple store to hold the data. This gave us, with a minimum of work, the possibility to query data using SPARQL. A SPARQL endpoint conforming to the SPARQL Protocol for RDF (Clark, Feigenbaum, Torres, 2008) was implemented to allow queries over HTTP.

4. Types of Resources Described

There are a number of types of resources that needs to be described or made available to reflect the current state of a library catalogue, e.g books, authors, subjects (for controlled vocabularies and thesauri), organizations, links between them, etc. To make the library catalogue available to systems outside the library community, the resources should be described using common vocabularies. We used Dublin Core for bibliographic data, FOAF[36] for persons and organizations, and SKOS[37] for controlled vocabularies. These are all widely used and understood standards. An example graph is displayed in FIG. 2. See Appendix A for example records.

It is important to point out that it is possible to deliver multiple formats in parallel, so catering to the world outside the library community does not exclude systems aware of library standards. As described in 3.1 RDF Server Wrapper adding support for Bibliontology, MODS, MarcOnt or any other standard is easy, it is, however, not the subject of this paper.

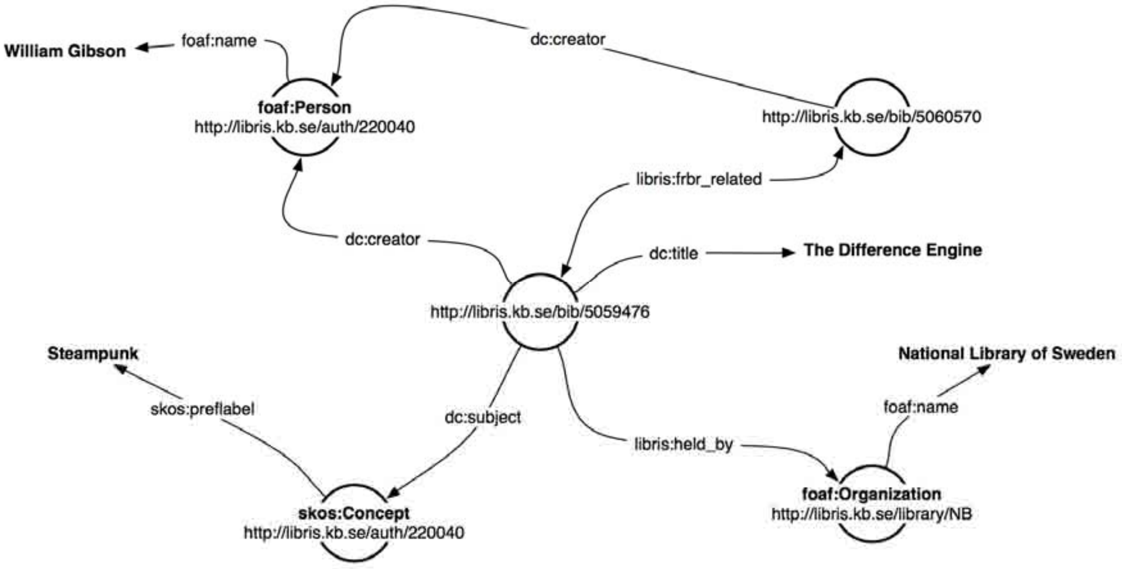


FIG. 2. Partial graph for the book "The Difference Engine"

5. FRBR

The Functional Requirements for Bibliographic Records (IFLA Section on Cataloguing, 1998) has been around for a decade, much has been written about it, though actual implementations are few. One hurdle to overcome is the shifting quality of the records due to continually changing practices. However, the idea of grouping or linking records being part of the same work is an appealing and technically viable one.

Every record in the LIBRIS database gets assigned one or more FRBR-keys, these keys are the normalized concatenations of an author and the original title. The process is repeated for each author and title. For example, the book "The Difference Engine" by William Gibson and Bruce

[35] Jena Semantic Web Framework - http://jena.sourceforge.net/
[36] Friend of a Friend - http://www.foaf-project.org/
[37] Simple Knowledge Organization System - http://www.w3.org/2004/02/skos/

Sterling has two keys: “GIBSON WILLIAM 1948 THE DIFFERENCE ENGINE”, and “STERLING BRUCE 1954 THE DIFFERENCE ENGINE”. Links are then created between records with the same key.

This is similar to the approach of Styles et al. (2008) where the MD5 checksum of the name of the author and the title of the work are used as an identifier.

However, an important distinction compared to Styles et al. is that these keys are transient; they are never used as identifiers, only to create the links between records of the same work. This way, when an author dies, changes his/her name, etc. the links remain the same even though the keys change. There is therefore no need to keep track of changes since no identifier has been published. Another advantage is that works with more than one author is handled automatically, as well as records containing more than one work.

The LIBRIS database also contains actual work records in the form of name+title authority records. These are linked to their respective bibliographic records. The sheer amount of bibliographic records prohibits manual creation of these for the whole database, nevertheless these links are included in the RDF.

6. Links to External Resources

Linking to external resources gives the client a way of finding more information about a given resource. As a proof-of-concept the LIBRIS database contains a handful of links from authority records to DBpedia and Wikipedia. See Appendix A for an example.

We have also experimented with user annotation using the annotea ontology. Since the URIs used to identify records/resources are available outside the ILS, attaching data, such as user reviews, to them is easy and non-intrusive.

7. Conclusion

Although there are a number of ontologies available to describe bibliographic data, the data contained in library systems are not generally available. The access mechanisms described in Linked Data need to be implemented for libraries to truly be “part of the semantic web”.

SPARQL shows real promise when it comes to mining the bibliographic data for information due to it’s linked nature.

Planned next steps include using SPARQL for automatic creation of work records, implementing a richer description of bibliographic and authority records and loading more external data into the triple store. We are closely following the work of the DCMI/RDA Task Group[38].

We are currently exploring the possibility of making parts of this work available as Open Source. More information will be available at http://libris.kb.se/semweb.

[38] http://dublincore.org/dcmirdataskgroup/

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. (2001). The Semantic Web. *Scientific American, 284*, 34-43.

Berners-Lee, Tim. (2006). *Linked data*. Retrieved April 12, 2008, from http://www.w3.org/DesignIssues/LinkedData.html.

Prud’hommeaux, Eric, and Andy Seaborn. (2008). *SPARQL Query Language for RDF*. Retrieved April 12, 2008 from http://www.w3.org/TR/rdf-sparql-query/.

Clark, James. (1999). *XSL Transformations (XSLT)*. Retrieved April 13, 2008 from http://www.w3.org/TR/xslt.

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional Requirements for Bibliographic Records*. Retrieved April 13, 2008 from http://www.ifla.org/VII/s13/frbr/frbr.pdf.

Styles, Rob, Danny Ayers, and Nadeem Shabir. (2008). Semantic MARC, MARC21 and the Semantic Web. *WWW 2008 17th International World Wide Web Conference*.

Sauermann, Leo, and Richard Cyganiak. (2008). *Cool URIs for the Semantic Web*. Retrieved April 13, 2008 from http://www.w3.org/TR/cooluris/.

Appendix A. - Examples of HTTP Requests and Responses

The following are HTTP traces of requests for bibliographic and authority records.

1. Bibliographic record - request, redirect and response

```
GET /resource/bib/5059476
Host: libris.kb.se
Accept: text/rdf+n3
----------------------------------------
HTTP/1.1 303 See Other
Location: http://libris.kb.se/data/bib/5059476
----------------------------------------
GET /data/bib/5059476
Host: libris.kb.se
Accept: text/rdf+n3
----------------------------------------
HTTP/1.1 200 OK
Content-Type: text/rdf+n3

@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix libris: <http://libris.kb.se/experimental/> .
@prefix annotea: <http://www.w3.org/2000/10/annotation-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# RDF in Turtle/N3 created for the bibliographic record 5059476
<http://libris.kb.se/resource/bib/5059476>
       foaf:page <http://libris.kb.se/bib/5059476>;
       rdfs:isDefinedBy <http://libris.kb.se/data/bib/5059476>;

       # short bibliographic description
       dc:title          "The difference engine";
       dc:creator  "Gibson, William, 1948-";
       dc:creator  "Sterling, Bruce, 1954-";
       dc:subject  "Steampunk";
       dc:identifier <URN:ISBN:0-575-04762-3>;
       ...

       # links to authors with authority records
       dc:creator  <http://libris.kb.se/resource/auth/220040>;
       dc:creator  <http://libris.kb.se/resource/auth/307779>;

       # links to subjects with authority records
       dc:subject  <http://libris.kb.se/resource/auth/308073>;
       dc:subject  <http://libris.kb.se/resource/auth/308074>;

       # links to other editions of the same work
       libris:frbr_related <http://libris.kb.se/resource/bib/5060570>;

       # user annotations
       annotea:hasAnnotation <http://libris.kb.se/resource/annotation/123>;

       # book is held by the following libraries
       libris:held_by    <http://libris.kb.se/resource/library/Sk>;
       libris:held_by    <http://libris.kb.se/resource/library/Vvt> .
```

2. Authority Record for Author William Gibson - Request and Response

```
GET /data/auth/220040
Host: libris.kb.se
Accept: text/rdf+n3
----------------------------------------
HTTP/1.1 200 OK
Content-Type: text/rdf+n3

@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dbpedia: <http://dbpedia.org/property/> .
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# RDF in Turtle/N3 created for the authority record 220040
<http://libris.kb.se/resource/auth/220040>
       rdfs:isDefinedBy <http://libris.kb.se/data/auth/220040>;

 # type of authority record
 rdf:type<http://xmlns.com/foaf/0.1/Person> ;

 # description
 foaf:name "Gibson, William, 1948-" ;
 foaf:name "William Gibson" ;
 ...

 # links to external resources
 owl:sameAs  <http://dbpedia.org/data/William_Gibson> ;
 rdfs:seeAlso <http://en.wikipedia.org/wiki/William_Gibson> .

# links to books by this author
<http://libris.kb.se/resource/bib/2716178>                            dc:creator
<http://libris.kb.se/resource/auth/220040> .
<http://libris.kb.se/resource/bib/2793076>                            dc:creator
<http://libris.kb.se/resource/auth/220040> .
<http://libris.kb.se/resource/bib/4465470>                            dc:creator
<http://libris.kb.se/resource/auth/220040> .
<http://libris.kb.se/resource/bib/4574314>                            dc:creator
<http://libris.kb.se/resource/auth/220040> .
...
```

3. Authority Record for the Subject Steampunk - Request and Response

```
GET /data/auth/308074
Host: libris.kb.se
Accept: text/rdf+n3
----------------------------------------
HTTP/1.1 200 OK
Content-Type: text/rdf+n3

@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

# RDF in Turtle/N3 created for the authority record 308074
<http://libris.kb.se/resource/auth/308074>
       rdfs:isDefinedBy <http://libris.kb.se/data/auth/308074>;

       # type of authority record
       rdf:type         skos:Concept ;

       # description
       skos:prefLabel  "Steampunk" ;
       skos:related    "Science fiction" ;
       skos:related    "Cyberpunk" ;

       # links to other subjects
       skos:related    <http://libris.kb.se/resource/auth/243892> ;
       skos:related    <http://libris.kb.se/resource/auth/142481> ;

       # links to external resources
       owl:sameAs <http://dbpedia.org/page/Steampunk> .

# links to books with this subject
<http://libris.kb.se/resource/bib/5059476>                            dc:subject
<http://libris.kb.se/resource/auth/308074> .
<http://libris.kb.se/resource/bib/5060570>                            dc:subject
<http://libris.kb.se/resource/auth/308074> .
```

Project Reports

Session 2:

Metadata Scheme Design, Application, and Use

The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment

Hollie C. White
University of North Carolina at Chapel Hill, USA
hcwhite1@email.unc.edu

Sarah Carrier
University of North Carolina at Chapel Hill, USA
scarrier@email.unc.edu

Abbey Thompson
University of North Carolina at Chapel Hill, USA
abbeyth@email.unc.edu

Jane Greenberg
University of North Carolina at Chapel Hill, USA
janeg@email.unc.edu

Ryan Scherle
National Evolutionary Synthesis Center, USA
rscherle@nescent.org

Abstract

This report presents recent metadata developments for Dryad, a digital repository hosting datasets underlying publications in the field of evolutionary biology. We review our efforts to bring the Dryad application profile into conformance with the Singapore Framework and discuss practical issues underlying the application profile implementation in a DSpace environment. The report concludes by outlining the next steps planned as Dryad moves into the next phase of development.

Keywords: Dryad; application profile; Singapore Framework; metadata generation; DSpace

1. Introduction

The Dryad repository[39] is a partnership between the National Evolutionary Synthesis Center (NESCent)[40] and the School of Information and Library Science, Metadata Research Center (SILS/MRC)[41] at the University of North Carolina at Chapel Hill. The repository hosts data supporting published research in the field of evolutionary biology. Dryad is currently working collaboratively with ten leading journals that publish evolutionary biology research, including *Evolution*, *The American Naturalist*, and *Ecology*. These journals have agreed to integrate their submission systems with Dryad in the near future, eventually creating a seamless publication process from author to journal to Dryad data deposition.

Two goals informing Dryad's current metadata activities include:

1. Dryad's need to be interoperable with other data repositories used by evolutionary biologists; and
2. Dryad's need for a sustainable information infrastructure.

The first goal has inspired our development of the Dryad application profile, version 1.0; and the second goal has led to Dryad's adoption of DSpace software and technology. Current metadata activities for the Dryad development team include revising the project's application profile so that it is compliant with the Singapore Framework. The Singapore Framework is a model that was released at the 2007 Dublin Core conference approximately a year after our team created the DRIADE application profile, version 1.0 (renamed Dryad application profile, ver.1.0) (Carrier, et al, 2007). Ongoing Dryad metadata work also includes evaluating the effectiveness of

[39] Note that in some previous publications Dryad is referred to as DRIADE.
[40] http://www.nescent.org
[41] http://ils.unc.edu/mrc/

our revised application profile and integrating it into a DSpace environment. This report reviews these two metadata focused activities, and highlights recent accomplishments and challenges.

2. Dryad's Application Profile

Dryad’s metadata application profile, ver.1.0, has two modules; one module describes data objects, and the other module describes the associating publication. We developed the application profile to support basic resource and data discovery, with the goal of being interoperable with other data repositories used by evolutionary biologists. The application profile is designed to automatically capture as much metadata as possible during publication and data deposition processing. The application profile incorporates elements from the following established metadata schemes: Dublin Core, Darwin Core, Data Documentation Initiative (DDI), Ecological Metadata Language (EML), and PREservation Metadata Implementation Strategies (PREMIS). The Dryad application profile, ver. 1.0, supports Dryad’s phase one functionalities that were established in a stakeholders’ workshop in December 2006[42]. These functionalities include the capturing, basic preservation, and simple retrieval of datasets and metadata for associated publications. In the future, metadata elements from other metadata schemes will be needed for projected features. Dryad’s phased development and corresponding functionalities are summarized in Table 1.

TABLE 1: Dryad Phased Implementation.

Phased Development/Implementation	Repository Functionalities
Phase One	• basic data/metadata storage • simple submission system
Phase Two	• integrate data deposition with publication • one-stop-deposition • data automatically and manually curated to ensure validity • automated metadata generation

3. DSpace and Dryad’s Metadata Architecture

DSpace is a software package for digital repository systems[43]. DSpace provides basic services to deposit, store, search, and retrieve digital content, but it was designed for a particular use case (storing publications, organized according to a university hierarchy), and significant modifications will be required to make DSpace suit the needs of Dryad users. Although the DSpace infrastructure has been adopted by many repositories, research on the integration of application profiles, especially those complying with the Singapore Framework, is still limited. Implementing the first iteration of the Dryad application profile in DSpace is allowing us to test the application profile, as well as evaluate the long-term applicability of DSpace for Dryad’s needs.

DSpace was chosen due to its adaptability and support of Dublin Core metadata, as well as the DSpace community’s support for enhancing metadata functionality, as evidenced by developments such as the SKOS module. Although most DSpace functionality revolves around qualified Dublin Core metadata, the software collects additional metadata that can be used to fill in details of the application profile, including qualifiers associated with elements drawn from

[42] https://www.nescent.org/wg_digitaldata/Dec_5_Workshop_Minutes

[43] http://www.dspace.org/

other metadata schemes. Metadata fields not native to DSpace are configured as custom fields, which can be stored, searched, and displayed in the same manner as the native fields.

A major advantage of DSpace is its system for managing user accounts, which can be adapted for the eventual Dryad functionality of allowing end-users to submit content and create basic metadata. However, the default workflow for submitting content and generating metadata in DSpace is entirely too long and awkward for end-users, and is further complicated by the needs of the Dryad metadata model. A more configurable submission system is included in the recently released DSpace 1.5, but significant work will still be required to allow users to submit content without difficulty.

One drawback of the DSpace model is that metadata with hierarchical information (e.g., MODS) are not supported by the core repository. Hierarchical information, which is necessary for tracking data such as contact information for multiple authors of a publication, must be stored in an extra file (bitstream) attached to the object, and modifications must be made to the default DSpace functionality if any of this information is to be used beyond simple display.

Another difficulty of using DSpace is the lack of a configurable access control system, a critical feature for Dryad. One requirement of Dryad is to collect and store publications to facilitate automatic metadata generation, while simultaneously shielding these publications from end-users. Some of the content stored in Dryad will need to be placed under embargo. While others have implemented these features in DSpace, the core distribution does not include them. Modifications to the core DSpace code must be kept to a minimum if we are to take advantage of future upgrades. Therefore, it will be challenging to optimize Dryad for users and metadata creators while minimizing deviation from the core DSpace platform.

4. Progressing toward Singapore Framework Compliance

The Singapore Framework provides a model for the structure of Dublin Core application profiles (Nilsson, Baker, & Johnston, 2008). Conformance with the Singapore Framework includes the benefits of consistency, long-term quality control, and interoperability with other metadata structures. A significant effort over the last few months has been to bring the Dryad application profile, ver. 1, which is based largely on Dublin Core, in line with the Singapore Framework. Reasons for this step include the benefits noted by Nilsson, et al. (2008), as well as, our goal to comply and interoperate with Semantic Web standards.

All five Singapore Framework components have been examined for the Dryad metadata schema adaptation (Carrier, 2008). The five components include the following: 1. Functional requirements; 2. Domain model; 3. Description Set Profile; 4. Usage guidelines; and 5. Encoding syntax guidelines. With the exception of the optional encoding syntax guidelines, the other four components have been deemed appropriate for the Dryad's application profile revision. The Scholarly Works Application Profile (SWAP)[44] is a key example of an application profile in conformance with the Singapore Framework, and provides a model for the Dryad description. The results of the initial restructuring can be found online as part of the repository project wiki.[45]

[44] http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

[45] https://www.nescent.org/wg_digitaldata/Level_One_Application_Profile

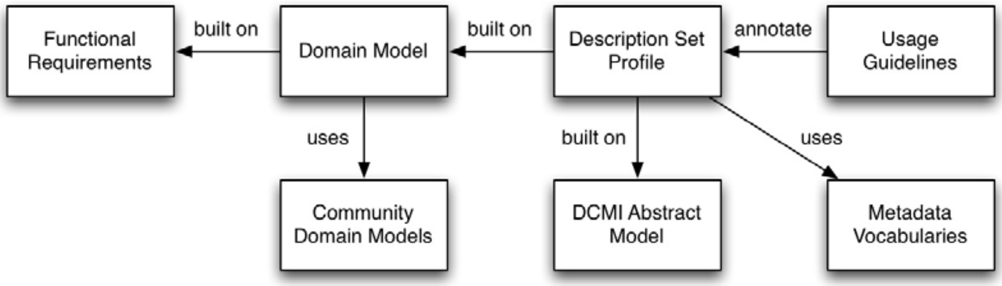


FIG. 1. Graphical representation of the Dryad Application Profile in the Singapore Framework model.

Addressing the Singapore Framework's first mandatory component, Dryad's functional requirements are based on project system requirement specifications. Using the SWAP example as a model, the Dryad's functional requirements (summarized in Table 1) address scope, stakeholders and designated community, requirements gathering, and functional requirements. Dryad's functional requirements include supporting the following operations: 1. resource discovery and use; 2. data interoperability; 3. computer-aided metadata generation and augmentation; 4. linking publications and underlying datasets; 5. data and metadata quality control; and 6. Data security. The designated community for the Dryad application profile includes researchers in the field of evolutionary biology who are generating data and reusing data for their own projects and scientists searching for datasets that are applicable to their own research. Stakeholders are evolutionary biologists, journal publishers in the field of evolutionary biology, professional societies in evolutionary biology, and NESCent. The methodology employed to gather system requirements involved assessing the needs and goals of individuals and groups identified as stakeholders and community members through a workshop held in December 2006 at NESCent in Durham, North Carolina, and more recently an ongoing use case study. Full details about the application profile functional requirements have been added to the Dryad project wiki46.The second mandatory component of the Singapore Framework is the domain model. Unlike the SWAP example, the Dryad application profile is "data-centric" rather than document- or publication-centric. Dryad's application profile, ver. 1.0, accommodates a single publication or article with published data from one or more datasets. This relationship is represented in Figure 2.

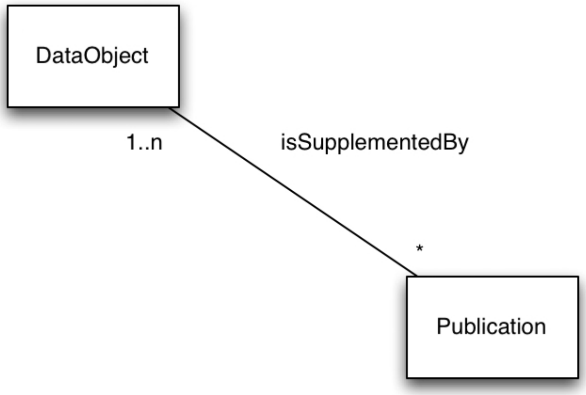


FIG. 2. Dryad Singapore Framework Domain Model

[46] https://www.nescent.org/wg_digitaldata/Level_One_Application_Profile#Functional_requirements

The third mandatory component, the *Description Set Profile (DSP)* is proving to be the most challenging aspect of the application profile revision process. As previously mentioned, the Dryad application profile is based largely on Dublin Core, but also incorporates elements from domain-specific namespaces such as PRISM, DDI, EML, and DarwinCore. None of the namespaces, except Dublin Core, are currently represented in RDF and cannot be included in the DSP. The Dryad development team has been discussing whether or not to declare unique elements for Dryad use in order to complete the Description Set Profile. Despite this challenge, the first draft of the Dryad DSP, which only includes Dublin Core elements, is available for viewing[47].

The fourth component, which is optional, is the *usage guidelines*, which have been collaboratively developed by Dryad team members and also appear online. The Dryad usage guidelines provide descriptions of each element and details regarding use[48]. Additionally, the guidelines also elaborate upon the constraints defined by the DSP.

5. Challenges and Future Work

The application profile revisions undertaken to comply with the Singapore Framework has strengthened the overall metadata architecture of the Dryad repository. It has also helped the project team identify key challenges, such as limitations in the current state of citation metadata, and the project's need to encode rights metadata. Furthermore, it has aided the Dryad development team in identifying metadata issues, and clarifying those issues that require administrative or policy decision, prior to determining the appropriate metadata element or value.

The most pressing issue facing the Dryad team is to determine how or if elements from non-Dublin Core namespaces should be included in the Dryad DSP and how the elements will be represented during DSpace implementation. The inclination is to use what has already been determined by a community to be useful, and furthermore to take advantage of the work and documentation already available from other initiatives; however, the issues with interoperability remain unavoidable at this time. Therefore, the Dryad team may choose to declare unique elements for the repository project.

The benefits of moving forward in line with the Singapore Framework are critical to the long-term success of Dryad and its ability to take advantage of metadata to improve system performance. The ongoing revision of the Dryad application profile, ver. 1, will result in the release and publication of the Dryad application profile, ver. 2.0. As part of our application profile development work, we are also taking into account selected functionalities of Dryad's phase two (Table 1). Additional ongoing activities include revising Dryad's interface for entering metadata and streamlining the metadata creation and submission process to support author-depositors. As Dryad evolves, we are anticipating that the recent release of DSpace 1.5 will impact the amount of work the project is able to complete with respect to specific metadata goals and other desired functionalities. In conclusion, Dryad's metadata structure is evolving, and will be revised over time, taking into consideration Semantic Web standards and innovations that support the overall goals of Dryad.

Acknowledgements

This work is supported by National Science Foundation Grant # EF-0423641. We would like to acknowledge contributions by the Dryad team members Hilmar Lapp and Todd Vision of NESCent.

[47] http://www.ils.unc.edu/~scarrier/dryad/DSPLevelOneAppProf.xml
[48] https://www.nescent.org/wg_digitaldata/Dryad_Level_One_Cataloging_Guidelines

References

Carrier, Sarah. (2008). *The Dryad Repository Application Profile: Process, development, and refinement*. Retrieved April 24, 2008 from http://hdl.handle.net/1901/534.

Carrier, Sarah, Jed Dube, and Jane Greenberg. (2007). The DRIADE project: Phased application profile development in support of open science. *International Conference on Dublin Core and Metadata Applications, Singapore, 2007.*

Nilsson, Mikael, Thomas Baker, and Pete Johnston. (2008). *The Singapore Framework for Dublin Core Application Profiles.* Retrieved April 10, 2008, from http://dublincore.org/documents/singapore-framework/.

Applying DCMI Elements to Digital Images and Text in the Archimedes Palimpsest Program

Michael B. Toth
Walters Art Museum –
R.B. Toth Associates, USA
mbt.rbtoth@gmail.com

Doug Emery
Walters Art Museum –
Emery IT, USA
doug@emeryit.com

Abstract

The digitized version of the only extant copy of Archimedes' key mathematical and scientific works contains over 6,500 images and 130 pages of transcriptions. Metadata is essential for managing, integrating and accessing these digital resources in the Web 2.0 environment. The Dublin Core Metadata Element Set meets many of our needs. It offers the needed flexibility and applicability to a variety of data sets containing different texts and images in a dynamic technical environment. The program team has continued to refine its data dictionary and elements based on the Dublin Core standard and feedback from the Dublin Core community since the 2006 Dublin Core Conference. This presentation cites the application and utility of the DCMI Standards during the final phase of this decade-long program. Since the 2006 conference, the amount of data has grown tenfold with new imaging techniques. Use of the DCMI Standards for integration across digital images and transcriptions will allow the hosting and integration of this data set and other cultural works across service providers, libraries and cultural institutions.

Keywords: Dublin Core; metadata standards; archiving; imaging; manuscript; Archimedes Palimpsest; cultural heritage; digital library

1. Introduction

Effective metadata standards are required to efficiently handle the large amounts of data collected in imaging and scholarly studies of the earliest known copy of Archimedes' work. The *Dublin Core Metadata Element Set* is being utilized to provide key identification information, with additional metadata extensions to ensure the imaging and scholarly information can be readily integrated in a Web 2.0 environment. Applying the Dublin Core Metadata Initiative (DCMI) Metadata Element Set and additional elements from the DCTerms namespace to a variety of images containing different texts in a rapidly changing technology environment has posed a unique set of challenges. These challenges include linking together and integrating data from different sources and formats: Digital images from advanced cameras in numerous spectral bands, and digitally encoded texts in varied fonts from a team of scholars. With increased focus on data management and explosive growth in data with advanced imaging techniques, the application of the *DCMI Metadata Element Set* provides a robust data set that will meet worldwide metadata standards.

2. Archimedes Palimpsest Program

The Archimedes Palimpsest Program is a 10-year effort to produce digital images of Archimedes' text as originally written on parchment in the latter half of the tenth century. In the early thirteenth century, this text was scraped off and overwritten, or "palimpsested," to create a prayer book. A team of scientists and scholars has been digitally imaging and studying the 174 parchment leaves that currently make up the Archimedes Palimpsest. Since the 2006 Dublin Core Conference the program has developed new imaging techniques that have yielded over two terabytes of data. This includes images of the only copies of Archimedes treatises *The Method* and *Stomachion*; the only copy in Greek of *On Floating Bodies;* and copies of the *Equilibrium of Planes*, *Spiral Lines*, *The Measurement of the Circle*, and *Sphere and Cylinder*. Imaging has also revealed ten leaves of text by the fourth century B.C. Greek orator Hyperides; six leaves of commentaries on Aristotle; four liturgical leaves; and twelve leaves from two unidentified books.

2.1. Imaging

At the time of the 2006 Dublin Core Conference, the imaging team had imaged the entire palimpsest with three spectral bands of light, yielding three images for each leaf and processed "pseudocolor" images in what was then considered to be a large data set of about 240 Gigabytes of data. Since the 2006 conference, the imaging team developed new imaging techniques to yield more information with more advanced cameras and lighting in 12 spectral bands. These yielded 16 images of each leaf and more refined processed images with a total of about 2,400 GB of data. (See Figure 1) Managing all this data required careful metadata logging and data management based on the *Dublin Core Metadata Element Set*.

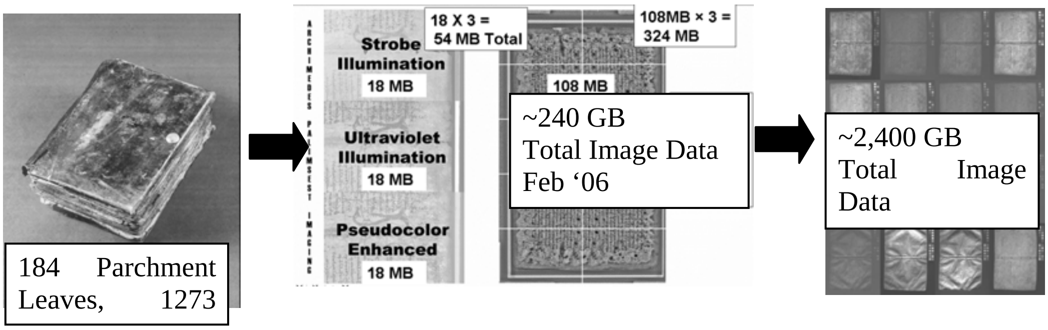


Fig. 1. Archimedes Palimpsest Data Growth © Images Copyright Owner of the Archimedes Palimpsest.

The Archimedes Palimpsest team also created images of key leaves at the Stanford Synchrotron Radiation Laboratory using X-ray fluorescence. This required an extensive range of metadata extensions to capture the broad range of metadata on energy levels and system parameters. They also imaged original prints of photographs of the Archimedes Palimpsest taken almost 100 years earlier at the direction of John Ludwig Heiberg in Constantinople, and photographs of one leaf taken in Chicago in the 1930's. These images of the photographs offered standardized images of text that has since been lost, and one leaf that has been lost in its entirety.

2.2. Metadata

With 6,797 digital images and 130 pages of transcriptions of the Archimedes Palimpsest, metadata has proved to be essential for 1) accessing images and integrating spectral bands for digital processing and enhancement, 2) managing transcriptions from those images for study by scholars around the world, and 3) linking and integrating the images and the transcriptions. This work required extensive identification metadata to ensure the data was manageable, as well as spatial metadata to line up and register the various images.

The *Dublin Core Metadata Element Set* offers the key identification elements required for image storage, management and retrieval, with additional spatial and spectral information added as extensions. The Archimedes Palimpsest Metadata Standard incorporates the DCMI Standards with six types of metadata elements:

1. Identification Information
2. Spatial Data Reference Information
3. Imaging and Spectral Data Reference Information
4. Data Type Information
5. Data Content Information
6. Metadata Reference Information

The “Identification,” “Data Type” and “Data Content” metadata elements incorporate the *Dublin Core Metadata Element Set.* The “Spatial Data Reference” and “Imaging and Spectral Data Reference” elements are extensions to the DCMI Standards, using metadata elements detailed in the Federal Geographic Data Committee *Content Standard for Digital Geospatial Metadata*. The standard is hosted on the www.archimedespalimpsest.org website.

In keeping with the project goals of long term data and metadata accessibilty, the program follows the DCMI’s principble of simplicity. We have created records that are machine-readable with very little effort and easily intelligible by a human reader. Each image metadata record is a series of simple name-value pairs, employing Dublin Core and project-specific metadata elements (See Table 1).

TABLE 1. Image Metadata Elements

Identifier	60000
Date	2008-03-03T08:20:56-05:00
Creator	Christens-Barry, Bill
Creator	Easton, Roger
Creator	Knox, Keith
Subject	Euchologion Image
Subject	Archimedes Palimpsest Image
Subject	Palimpsest Image
Subject	Multispectral Image
Subject	Digital Image
Subject	Greek Manuscript Image
Subject	Byzantine Manuscript Image
Subject	Private Collection
Publisher	Owner of the Archimedes Palimpsest
Contributor	Noel, Will
Contributor	Toth, Michael
Contributor	Auer, Kevin
Contributor	Emery, Doug
Contributor	Gerry, Kate
Contributor	Potter, Daniel
Contributor	Quandt, Abigail
Contributor	Tabritha, Ariel
Contributor	Tilghman, Ben
Contributor	Stokes, John R.
Type	Image
Source	Processed from image with Identifier 15380,0000100r_Arch53v_ Sinar_LED445 _01_raw.tif
Source	Processed from image with Identifier 15383, 0000-100r_Arch53v_Sinar_LED 530_01_raw.tif
Source	Processed from image with Identifier 15386, 0000-100r_Arch53v_Sinar_LED 625_01_raw.tif
Coverage	Walters Art Museum
Coverage	2007-08-06 to 2007-08-26
Coverage	Baltimore, MD
Coverage	USA
license	http://creativecommons.org/licenses/by/3.0/legalcode
license	Licensed for use under Creative Commons Attribution 3.0 Unported
accessRights	Copies of any articles published must be sent to William Noel, Walters Art Museum, Baltimore, MD.
ID_File_Name	0000-100r_Arch53v_Sinar_true_pack8.tif

Since the 2006 Dublin Core Conference, individuals with standards experience in OCLC and other organizations have provided input on the best application of the Dublin Core Standard to the Archimedes Palimpsest Metadata Standard. A range of organizations have also provided guidance on the use of standards for archival purposes, including guidance for the best use of standards in the digital data set to ensure users years hence will have access to the actual versions of the standards used in creating the data set. Input from the Library of Congress, the British Library, NASA and Google proved fruitful in defining the application of standards not only to the Web 2.0 environment, but the range of possible digital environments possible in decades to come.

2.3. Transcriptions

Scholars have been transcribing the Greek text since the initial digital imaging, revealing new information about the origins of mathematical theories and science. The integration of these scholarly transcriptions in digital form with the digital images has taken on greater impetus since the 2006 Dublin Core Conference, with the digital tagging and encoding of text in various forms and formats, including handwritten, MSWord Symbol font, and various other custom fonts. A team of scholars and students is encoding the transcribed text into XML tagged Unicode following the Text Encoding Initiative standards (See Figure2).

```
<seg TEIform="seg" n="17v1" part="N" type="folio">
   <seg TEIform="seg" n="1" part="N" type="line">
      <supplied TEIform="supplied" reason="lost">
         <expan TEIform="expan">ὅτι</expan> τὸ ΦΑ</supplied> μέγε<supplied
         TEIform="supplied" reason="lost">θος</supplied> τῶι βάρει πρὸς </seg>
   <seg TEIform="seg" n="2" part="N" type="line">
      <supplied TEIform="supplied" reason="lost">τὸ ὑγρ</supplied>ὸν τὸ ἰσόογκον
      τοῦτον ἔχει</seg>
   <seg TEIform="seg" n="3" part="N" type="line">τὸν λόγον, ὃν τὸ Α <expan
          TEIform="expan">πρὸς</expan> τὸ Φ<unclear TEIform="unclear"
          >Α</unclear>.</seg>
</seg>
```

FIG. 2. XRF Tagged Transcriptions

Header information is provided for each folio in the encoded text, with cross-walked Dublin Core Identification and Data Content metadata elements mapped to the TEI format (Figure 3). These encoded texts are then hosted with the images, with the Dublin Core elements providing a common structure for image and transcription metadata.

```
<teiHeader>
   <fileDesc>
      <titleStmt>
          <title>Transcription of fols. 17v-16r of the Archimedes Palimpsest
              (= Archimedes fol. 7v, On Floating Bodies)</title>
          <respStmt>
            <resp>Responsible for primary transcription (Dublin Core creator)</resp>
             <name>Reviel Netz</name>
          </respStmt>
          <respStmt>
             <resp>Contributor</resp>
             <name>Mike Toth</name>
          </respStmt>
<publicationStmt>
         <idno>5021</idno>
         <publisher>Owner of the Archimedes Palimpsest</publisher>
         <date>2008</date>
</publicationStmt>
   </fileDesc>
   <profileDesc>
      <langUsage>
        <language id="grc-c">accented ancient Greek in Unicode-C Greek
characters</language>
      </langUsage>
```

FIG 3. Sample Dublin Core Header Information in Encoded Transcription Headers.

3. Integrated Product

The images and transcriptions are linked through metadata in the Archimedes Palimpsest Data Product, enabling common searches, access and study. The standard use of the Dublin Core Metadata Element Set across the products of the image scientists and scholars enables linkage between these two disparate data sets for further study (Figure 2). Integrating metadata of various types tailored to meet a range of users' needs has proven critical to making integrated data available across domains and disciplines amidst ever changing technologies. Building on the Archimedes Palimpsest application, the DCMI Standards are being used to integrate hyperspectral imaging of the Waldseemuller 1507 Map at the Library of Congress. The DCMI Standards serve as the basis for information discovery in the Web 2.0 environment, and hopefully for decades to come in future formats and technologies. This information will advance the study of the original manuscript by individuals around the world with ubiquitous access via the Internet.

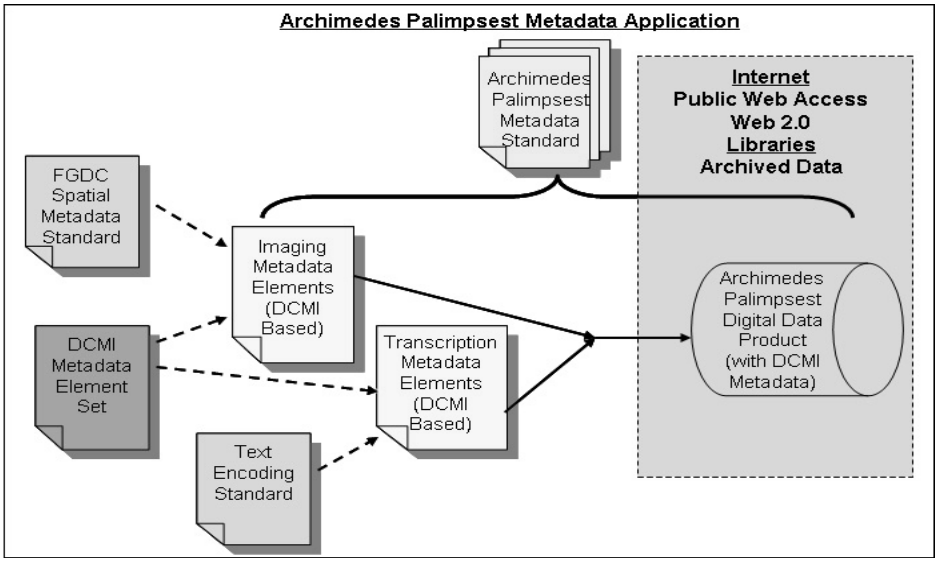

Archimedes Palimpsest Metadata Application
Archimedes Palimpsest Metadata Standard
Internet
Public Web Access
Web 2.0
Libraries
Archived Data
FGDC Spatial Metadata Standard
Imaging Metadata Elements (DCMI Based)
DCMI Metadata Element Set
Transcription Metadata Elements (DCMI Based)
Archimedes Palimpsest Digital Data Product (with DCMI Metadata)
Text Encoding Standard

Fig. 4. Archimedes Palimpsest Metadata Application Architecture.

Acknowledgements

The authors wish to thank the owner of the Archimedes Palimpsest for his support over the past decade and for making this important cultural data set available to the public. They would also like to thank their colleagues on the Archimedes Palimpsest Program team: Dr. William Noel for his leadership; Drs. Roger Easton, Keith Knox and William Christens-Barry for the imaging; Abigail Quandt and Jennifer Giaccai for the conservation; Reviel Netz and the team of scholars and students for the transcriptions; Carl Malamud for his data support; and OCLC and Bob Toth for metadata standards support. Archimedes Palimpsest images are ©Copyright the Owner of the Archimedes Palimpsest and used under Creative Commons 3.0 Attribution License.

References

Archimedes Palimpsest Program. (2006). *Archimedes Palimpsest Metadata Standard 1.0, Revision 5.* Baltimore, Maryland: Walters Art Museum (WAM).

Dublin Core Metadata Initiative. (2000-2008). *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Retrieved from http://dublincore.org/documents/dces/.

Dublin Core Metadata Initiative. (2000-2008). *DCMI Metadata Terms*. Retrieved from http://www.dublincore.org/documents/dcmi-terms/.

Federal Geospatial Data Committee. (2002). *Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata.* FGDC-STD-012-2002**.** Washington, DC: Federal Geospatial Data Committee.

Knox, Keith T., Roger L. Easton Jr., and William A. Christens-Barry. (2003). Multispectral imaging of the Archimedes Palimpsest. *2003 AMOS Conference.* Maui, Hawaii: Air Force Maui Optical & Supercomputing Site.

Netz, Reviel. (2000). The origin of mathematical physics: New light on an old question. *Physics Today*, 32-37.

Noel, Will, Roger L. Easton Jr., and Michael B. Toth. (2006). *The Archimedes Palimpsest.* California: Google Inc. Retrieved, March 7, 2008, from http://www.youtube.com/watch?v=S19Xyjxl4fI.

Toth, Michael B., William A. Christens-Barry, and Roger L. Easton Jr. (2006). Dublin Core based metadata supports the Archimedes Palimpsest Manuscript Imaging Program. *International Conference on Dublin Core and Metadata Applications, Colima, Mexico, October 3-6, 2006.*

WAM. (2008). *Archimedes - The Palimpsest*. Retrieved, March 20, 2008, from http://www.archimedespalimpsest.org/.

Assessing Descriptive Substance in Free-Text Collection-Level Metadata

Oksana Zavalina
University of Illinois at Urbana-Champaign, USA
zavalina@uiuc.edu

Carole L. Palmer
University of Illinois at Urbana-Champaign, USA
clpalmer@uiuc.edu

Amy S. Jackson
University of Illinois at Urbana-Champaign, USA
amyjacks@uiuc.edu

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

Abstract

Collection-level metadata has the potential to provide important information about the features and purpose of individual collections. This paper reports on a content analysis of collection records in an aggregation of cultural heritage collections. The findings show that the free-text *Description* field often provides more accurate and complete representation of subjects and object types than the specified fields. Properties such as importance, uniqueness, comprehensiveness, provenance, and creator are articulated, as well as other vital contextual information about the intentions of a collector and the value of a collection, as a whole, for scholarly users. The results demonstrate that the semantically rich free-text *Description* field is essential to understanding the context of collections in large aggregations and can serve as a source of data for enhancing and customizing controlled vocabularies.

Keywords: descriptive metadata; collection-level metadata; Dublin Core Collection Application Profile; federated digital collections; IMLS Digital Collections and Content project

1. Introduction and Background

It has long been recognized that contextual metadata is important for facilitating access to documents in archival collections (e.g., Bearman, 1992). More recently, digital collections have come to be understood as information seeking contexts (Allen & Sutton, 1993; Lee, 2000). As digital collections are aggregated into larger meta-collections, and grow in size and complexity, the need for a coherent contextual framework increases. Collection-level metadata can provide the necessary relational and contextual framework (Macgregor, 2003; Miller, 2000) through "unitary"[49] and "analytic"[50] descriptive approaches (Heaney, 2000).

Cultural heritage institutions have purposefully conceptualized and developed their digital collections in many ways, as "displays", "tours", "tools", "lessons", and to provide a record of cultural events (Palmer et al., 2006). However, in a large digital federation or aggregation, the purpose of the original, deliberately built collections becomes difficult to discern. Collection-level metadata has the potential to provide important information about features of a parent collection and why it might be of value to users. But the qualitative aspects of collections are difficult to describe in a systematic way, as they may embody a good deal of intellectual intent and tend to be highly complex and mutable.

This paper reports on the current phase of the Digital Collections and Content (DCC) project that is investigating how to represent collection context for scholarly use of large-scale, heterogeneous digital aggregations. The DCC provides integrated access to over 200 digital

[49] Defined as: "consists only of information about the collection as a whole."
[50] Defined as: "consists of information about the individual items within [a collection] and their content."

collections funded by the Institute of Museum and Library Services (IMLS), National Leadership Grant program, through a centralized collection registry and metadata repository. The DCC collection metadata schema used for the registry was adapted from a preliminary version of the Dublin Core Collection Description Application Profile (DC CDAP) and the UKOLN RSLP schema (Heaney, 2000). The information used to encode collection registry records is gathered directly from resource developers through a survey, with complementary information taken from collection websites and the descriptive text provided in the grant proposals submitted to IMLS. Once the initial record has been created, it is sent to the local collection administrator for review and editing. Needed updates, changes, and additions of information and links to related collections are made through the DCC collection record edit interface. The DCC project coordinator is responsible for final review and release of all collection records made accessible through the public interface.

Previous DCC reports have discussed the various ways that resource developers conceive of collections, the attributes they find most important in describing collections, and the different "cultures of description" evident among libraries, museums, archives, and historical societies (Knutson, Palmer, & Twidale, 2003; Palmer & Knutson, 2004). In addition, preliminary DCC usability studies suggested that collection and subcollection metadata help users ascertain features like uniqueness, authority, and representativeness of objects retrieved and can lessen confusion experienced searching large-scale federations (Foulonneau et al., 2005; Twidale & Urban, 2005). The analysis presented here builds on previous DCC work[51] to extend our understanding of the role of collection metadata and provide an empirical foundation for our ongoing analysis of item-level and collection-level metadata relationships (Renear et al., forthcoming).

2. Methods

The objectives of the study were to identify the range of substantive and purposeful information about collections available within the DCC Collection Registry, determine patterns of representation, and assess the adequacy of the DCC collection-level metadata schema[52] for representing the richness and diversity of collections in the aggregation. The results presented here are based on a systematic, manual analysis of 202 collection-level records. The free-text in the *Description* field was both qualitatively and quantitatively analyzed to identify types of information provided about a digital collection and the degree of agreement between information provided in the free-text *Description* field and relevant information found in other free-text and controlled vocabulary fields. Hereafter, we use the term "collection properties" to refer to the types of information identified in the collection records.[53]

3. Findings

Table 1 lists the properties found only in the *Description* field of the DCC collections record. The properties are subdivided into three groups. The first consists of three properties that are special claims about collections: Importance (e.g., "collection of the most important and influential 19th and early 20th century American cookbooks"), Uniqueness (e.g., "unique historical treasures from … archives, libraries, museums, and other repositories"), and Comprehensiveness (e.g., "a comprehensive and integrated collection of sources and resources on the history and topography of London"). These properties are of particular interest as the kind of

[51] Described in detail in our five-year report http://imlsdcc.grainger.uiuc.edu/docs/FinalReport_ResearchMethods.pdf

[52] Available at: http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp

[53] No predefined list of categories was used for analysis. The categories emerged from coding performed by two coders who are authors on this paper. A test of intercoder reliability showed 80.4% agreement in assigning the codes to specific cases.

self-assessed value commonly used to distinguish special collections. Although not prominent enough to include in the table, a related property, "Strength", appeared in three records.[54]

The second group contains two other common descriptive properties also not delineated in the DCC collection metadata schema: Creator of items in the collection (e.g., "The Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items currently in the collection") and Provenance (e.g., "in December 2002, the … Library acquired the Humphrey Winterton Collection of East African photographs"). Item Creator[55] and Provenance elements might serve an even greater number of DCC collections than those currently exploiting the *Description* field for these purposes. There are DCC collections related to single or multiple authors that could benefit from more formal representation of item creators. In this case, a new element would need to be specified, since the existing DC CDAP *Collector* element is designed to cover creator of the collection not creator of items in the digital collection. Also, a large number of the collections come from museums, and a smaller but substantial group from historical societies and archives. These institutions are likely to have conventions for documenting chain of custody. Here, the DC CDAP *Custodial History* element is a good model, since it covers the kind of provenance information found in our free-text metadata.

The third group contains Subject and Object. Formal elements do exist for these properties, but the analysis shows that the *Description* field provides extensive additional coverage (e.g., "broad range of topics, including ranching, mining, land grants, anti-Chinese movements, crime on the border, and governmental issues"; "souvenirs of all kinds, including plates, cups, vases, trays, bottles, sewing boxes and games").

TABLE 1. Collection properties unique to *Description* field.

Collection Property	***Number of collections***	***%***
GROUP 1		
Importance	20	10.1
Uniqueness	17	9.0
Comprehensiveness	6	3.0
GROUP 2		
Item Creator	78	39.4
Provenance	24	12.1
GROUP 3		
Subjects not represented in formal metadata elements	132	66.7
Objects not represented in formal metadata elements	37	18.7

TABLE 2. Other collection properties in *Description* field.

Collection Property	***Number of collections***	***%***
Subjects	181	91.4
Object types	149	75.3
Collection development policy	102	52.0
Collection title	103	52.0
Size	53	26.8
Audience	34	17.0
Navigation and functionality	32	16.2
Participating/contributing institutions	30	15.2
Funding sources	10	5.1

[54] See Johnston (2003) for discussion on inclusion of a Strength element in the Dublin Core Collection Description Application Profile.

[55] The DCC collection description metadata schema currently uses dc:creator element in a limited way to indicate a grant project responsible for creation of the digital collection, but does not include creators of items and collections.

Table 2 shows nine collection properties represented but not unique to the free-text *Description* field. The subject information in the *Description* field ranges from specific statements to subject keywords scattered throughout the text. In most cases (66.7%), the *Description* field provides more accurate and specific coverage than the fields intended for subject indexing*: Subjects*, *GEM Subjects*, *Geographic Coverage*, and *Time Period*. Fifty percent of the *Description* fields include indications of temporal coverage, ranging from specific dates and date ranges (e.g., 19th century) to known historical periods (e.g., World War I, California Golden Rush). Sixty percent of *Description* fields include indications of geographic coverage of varying granularity (e.g., "Austro-Hungarian Empire"; "Mayan city of Uxmal in Yucatan, Mexico and a Native American Mississippian site, Angel Mounds U.S.A.").

The *Description* field often lists additional, or more specific, types of objects than covered by the formal element, *Objects Represented*. Broad terms, such as "physical artifacts", are common, as are more specific terms, such as "lanterns, torches, banners". Formats and genres are also frequently specified, as with "leaflets", "songbooks", and "political cartoons". Object types and formats are sometimes conflated, even within the same sentence, in the *Description* field, as well as in *Objects Represented*. This lack of disambiguation between type and format is a known metadata quality problem in digital object description (see, for example, Jackson et al., 2008).

Over half of the *Description* fields contain evidence of collection development policies (e.g., "titles published between 1850 and 1950 were selected and ranked by teams of scholars for their great historical importance"). Some identify other locally accessible materials or plans for future collection development, a potentially significant aspect of collector intentionality: "it is planned to provide access to a complimentary collection of Richmond related Civil War period resources"; "lesson plans, activities and photo essays designed by teacher advisors and educational consultants will be added in the future". Others explicitly state a purpose: "support global efforts to conserve, study, and appreciate the diversity of palms".

While duplicative of the *Title* field, many titles found in the *Description* field (either full title or part of title) provide concise statements with subject-specific information, as well as information on the object types in a collection. Collection size statements in the *Description* field range from quantitative specifications (e.g., "209 cartoons, 12 Christmas cards, and 3 facsimiles of cartoons") to general orientations (e.g., "hundreds of personal letters, diaries, photos, and maps"). In 28% of the cases, the *Description* field is the only source of this important information. In 30% of the collection records the size data in the *Description* and *Size* fields do not match; these discrepancies seem to reflect, sometimes clearly, the difference between projected and actual size of the digital collection (e.g., "When finished, the collection guide will consist of well over 100,000 online stereoviews" in the *Description* field and "38254 Stereographic Photoprints" in the *Size* field).

Audience information, found in 17% of *Description* fields (e.g., "Alabama residents and students, researchers, and the general public"), often complements and clarifies controlled vocabulary values in the *Audience* field. For example, in a record where the *Audience* field lists "General public, K-12 students, undergraduate students, K-12 teachers and administrators, Scholars/researchers/graduate students", the *Description* field specifies "anthropologists, art historians, cultural studies scholars, historians, political scientists and sociologists".

Some aspects of navigation or functionality represented in the *Description* field are also found in the formal *Interaction with Collection* field of the same record (e.g., "accessible by date of issue or by keyword searching" in *Description* and "search, browse" in *Interaction with Collection*). In most cases, information in the two fields is complementary.

Institutions participating in the digitization project and contributing items to digitize (e.g., "project brings … together with the University to build a digital repository") and funding sources that helped support digital collections (e.g., "funds provided by the Institute of Museum and Library Services, under the federal Library Services and Technology Act") are also often acknowledged in *Description* fields.

4. Discussion and Conclusions

Our findings identify the various kinds of substantive descriptive information provided in the free-text *Description* element, much of which clearly enriches the collection-level records and provides important scholarly context for the collections within the DCC. There is consistent representation of subjects and object types that is more accurate in coverage and offers more detail than that represented in the other fields specified for those purposes. Moreover, "special claims" about a collection's importance, uniqueness, or comprehensiveness are not represented in any other way within the record and add vital qualitative and contextual information about the intentions of collectors and the role the collection plays in the larger universe of related content. Provenance and Item Creator properties are not accommodated in the current DCC collection metadata schema, but were strongly represented within the *Description* field. All of these data represent distinguishing features potentially of interest to scholarly and other research audiences.

Based on these findings, the first activity slated for collection record enhancement in the DCC is to align the DCC collection description schema with the DC CDAP, which was released after development of the DCC schema. The *Custodial History* field will accommodate some of the key information currently found only in the *Description* field. A newly defined field for creators of items in a collection and a specified field for special claims about collections are also under consideration. Moreover, the *Description* field is clearly a semantically-rich source from which to mine terms to develop a customized controlled vocabulary for use in the DCC and similar aggregations of cultural heritage digital materials. The research team is exploring how to enhance the current controlled vocabulary with frequently used terms and concepts used in the *Description* field. This terminology would be more representative of the language used by collection creators to explain the purpose and value of their content and would provide a more accurate record of the materials included in cultural heritage collections. The next step in our study of free-text collection-level metadata is a comparative analysis of collection records from sources other than the DCC, produced by libraries, museums, and archives. A broader understanding of the use of the *Description* field in various organizational contexts will be particularly meaningful as we continue to explore the general relationship between content and context and the ways in which collection-level description can complement item-level description.

Acknowledgements

This research was supported by a 2007 IMLS National Leadership Research and Demonstration grant (LG-06-07-0020-07). We also wish to thank our colleagues from the Metadata Roundtable for their helpful comments and suggestions on a preliminary draft of this paper.

References

Allen, Bryce L., and Brett Sutton. (1993). Exploring the intellectual organization of an interdisciplinary research institute. *College & Research Libraries, 54,* 499–515.

Bearman, David. (1992). Contexts of creation and dissemination as approaches to documents that move and speak. *Documents that Move and Speak: Audiovisual Archives in the New Information Age: Proceedings of a Symposium 30 April -3 May 1990, National Archivesof Canada,* (pp. 140-149).

Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, and Sarah L.Shreeves. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* (pp. 32-41).

Heaney, Michael. (2000). *An analytical model of collections and their catalogues.* Retrieved April 12, 2008, from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf.

Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8* (1).

Johnston, Pete. (2003). *Report from meeting of DC CD WG at DC-2003.* Retrieved April 12, 2008, from http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0310&L=DC-COLLECTIONS&D=0&I=-3&P=59.

Knutson, Ellen M., Carole L. Palmer, and Michael Twidale. (2003). Tracking metadata use for digital collections. *Proceedings of the International DCMI Metadata Conference and Workshop,* (pp. 243-244).

Lee, Hur-Li. (2000). What is a collection?. *Journal of the American Society for Information Science, 51*(12), 1106-1113.

Macgregor, George. (2003). Collection-level descriptions: Metadata of the future? *Library Review, 52*(6), 247-250.

Miller, Paul. (2000, September). Collected wisdom: some cross-domain issues of collection-level description. *D-Lib Magazine, 6*(9). Retrieved June 14, 2008, from http://www.dlib.org/dlib/september00/miller/09miller.html.

Palmer, Carole L., and Ellen M. Knutson. (2004). Metadata practices and implications for federated collections. *Proceedings of the 67th ASIS&T Annual Meeting,* (pp. 456-462).

Palmer, Carole L., Ellen M. Knutson, Michael Twidale, and Oksana Zavalina. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting,* (pp. 161-162).

Renear, Allen H., Richard J. Urban, Karen M. Wickett, Carole L. Palmer, and David Dubin. (forthcoming). Sustaining collection value: Managing collection/item metadata relationships. *Proceedings of the Digital Humanities Conference, 25-29 June 2008, Oulu, Finland.*

Twidale, Michael, and Richard J. Urban. (2005). *Usability analysis of the IMLS Digital Collection Registry*. Retrieved June 14, 2008, from http://imlsdcc.grainger.uiuc.edu/3YearReport/docs/UsabilityReport1.pdf.

Project Reports

Session 3

Vocabulary Integration and Interoperability

Building a Terminology Network for Search: The KoMoHe Project

Philipp Mayr
GESIS Social Science Information Centre
(GESIS-IZ), Bonn, Germany
Philipp.Mayr@gesis.org

Vivien Petras
GESIS Social Science Information Centre
(GESIS-IZ), Bonn, Germany
Vivien.Petras@gesis.org

Abstract

The paper reports about results on the GESIS-IZ project "Competence Center Modeling and Treatment of Semantic Heterogeneity" (KoMoHe). KoMoHe supervised a terminology mapping effort, in which 'cross-concordances' between major controlled vocabularies were organized, created and managed. In this paper we describe the establishment and implementation of cross-concordances for search in a digital library (DL).

Keywords: cross-concordances; terminology mapping; terminology service; subject searching

1. Project Background

Semantic integration seeks to connect different information systems through their subject metadata frameworks – insuring that distributed search over several information systems can still use the advanced subject access tools provided with the individual databases. Through the mapping of different subject terminologies, a 'semantic agreement' for the overall collection to be searched on is achieved. Terminology mapping – the mapping of words and phrases of one controlled vocabulary to the words and phrases of another – creates a semantic network between the information systems carrying the advantages of controlled subject metadata schemes into the distributed digital library world.

Terminology mappings could support distributed search in several ways. First and foremost, they should enable seamless search in databases with different subject metadata systems. Additionally, they can serve as tools for vocabulary expansion in general since they present a vocabulary network of equivalent, broader, narrower and related term relationships (see examples in TAB. 1). Thirdly, this vocabulary network of semantic mappings can also be used for query expansion and reformulation.

Starting point of the project was the multidisciplinary science portal vascoda[56] which merges structured, high-quality information collections from more than 40 providers in one search interface. A concept was needed that tackles the semantic heterogeneity between different controlled vocabularies (Hellweg et al., 2001, Krause, 2003).

In 2004, the German Federal Ministry for Education and Research funded a major terminology mapping initiative (KoMoHe project[57]) at the GESIS Social Science Information Centre in Bonn (GESIS-IZ), which found its conclusion at the end of 2007. One task of this terminology mapping initiative was to organize, create and manage 'cross-concordances' between major controlled vocabularies (thesauri, classification systems, subject heading lists) centered around the social sciences but quickly extending to other subject areas (see FIG. 1). The main objective of the project was to establish, implement and evaluate a terminology network for search in a typical DL environment.

In this paper, we describe the establishment and implementation of cross-concordances for search. A thorough information retrieval evaluation of several cross-concordances analyzing their effect on search was undertaken and is described in Mayr & Petras (2008 to appear).

[56] http://www.vascoda.de/

[57] http://www.gesis.org/en/research/information_technology/komohe.htm

2. Building a Cross-concordance Network

We define cross-concordances as intellectually (manually) created crosswalks that determine equivalence, hierarchy, and association relations between terms from two controlled vocabularies.

Typically, vocabularies will be related bilaterally, that is, a cross-concordance relating terms from vocabulary A to vocabulary B as well as a cross-concordance relating terms from vocabulary B to vocabulary A are established. Bilateral relations are not necessarily symmetrical. For example, the term ‘Computer’ in system A is mapped to the term ‘Information System’ in system B, but the same term ‘Information System’ in system B is mapped to another term ‘Data base’ in system A.

Cross-concordances are only one approach to treat semantic heterogeneity (compare Hellweg et al., 2001, Zeng & Chan, 2004).

Our approach allows the following 1:1 or 1:n relations:

- Equivalence (=) means identity, synonym, quasi-synonym
- Hierarchy (Broader terms <; narrower terms >)
- Association (^) for related terms
- An exception is the Null (0) relation, which means that a term can’t be mapped to another term (see mapping number 4 in TAB. 1).

In addition, every relation must be tagged with a relevance rating (high, medium, and low). The relevance rating is a secondary but weak instrument to adjust the quality of the relations. They are not used in our current implementations. In our approach it takes approximately 4 minutes to establish one mapping between two concepts. Table 1 presents typical unidirectional cross-concordances between two vocabularies A and B.

TABLE 1. Cross-concordance examples (unidirectional).

No	Vocabulary A	Relation	Vocabulary B	Description
1	hacker	=	hacking	Equivalence relationship
2	hacker	^+	computers + crime	2 association relations (^) to term combinations (+)
3	hacker	^+	internet + security	
4	isdn device	0		Concept can’t be mapped, term is too specific.
5	isdn	<	telecommunications	Narrower term relationship
6	documentation system	>	abstracting services	Broader term relationship

The mappings in the KoMoHe project involve all or major parts of the vocabularies. Vocabularies were analyzed in terms of topical and syntactical overlap before the mapping started. Term lists are precompiled and ready to map when they come to people who are mapping. Collaborative work on one mapping is possible, but more complicated to organize. All mappings are created by researchers or terminology experts. Essential for a successful mapping is an understanding of the meaning and semantics of the terms and the internal relations (structure) of the concerned vocabularies[58]. This includes syntactic checks of word stems but also semantic knowledge to look up synonyms and other related terms. See in this context Lauser et al. (2008, to be published) for an insight concerning intellectual and automatic mapping methodologies.

[58] Some of the same problems occur in the development of multilingual thesauri, which are detailed in IFLA (2005) and the ISO 5964 (1985) standard.

The mapping process is based on a set of practical rules and guidelines (see also Patel et al., 2005). During the mapping of the terms, all intra-thesaurus relations (including scope notes) are consulted. Recall and precision of the established relations have to be checked in the associated databases. This is especially important for combinations of terms (1:n relations). One-to-one (1:1) term relations are preferred. Word groups and relevance adjustments have to be made consistently.

In the end, the semantics of the mappings are reviewed by experts and samples are empirically tested for document recall and precision. Expert reviews focus especially on semantic correctness, consistency and relevance of equivalence relations which are our most important relationship type. Sampled mappings are cross-checked and assessed via queries against the controlled term field of the associated database.

More mapping examples can be found in Mayr & Walter (2008).

To date, 25 controlled vocabularies from 11 disciplines and 3 languages (German, English and Russian) have been connected with vocabulary sizes ranging from 1,000 – 17,000 terms per vocabulary (see the project website for more details). More than 513,000 relations were generated in 64 crosswalks. Figure 1 depicts the established network of cross-concordances by discipline.

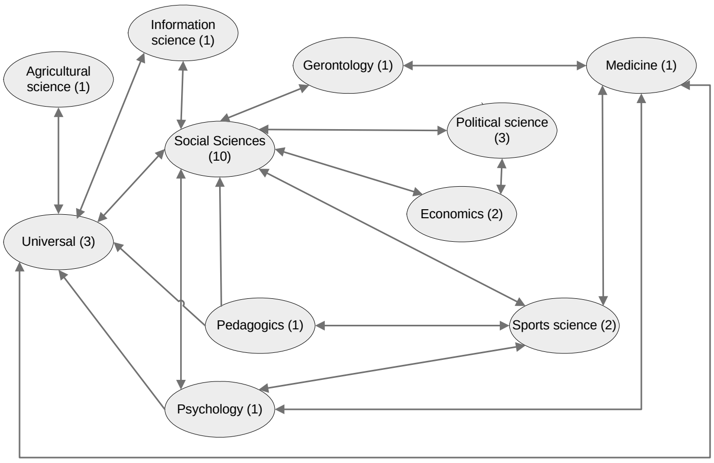


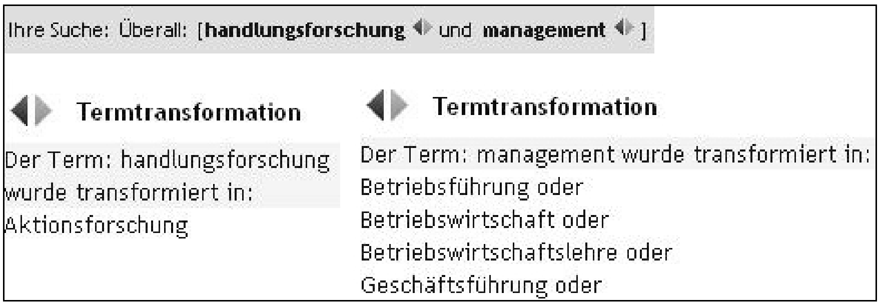
FIG. 1. Network of terminology mappings in the KoMoHe project. The numbers in brackets contain the number of mapped controlled vocabularies in a discipline.

3. Implementing Cross-concordances for Search

A relational database was created to store the cross-concordances for later use. It was found that the relational structure is able to capture the number of different controlled vocabularies, terms, term combinations, and relationships appropriately. The vocabularies and terms are represented in list form, independent from each other and without attention to the syndetic structure of the involved vocabularies. Orthography and capitalization of controlled vocabulary terms were normalized. Term combinations (i.e. computers + crime as related combination for the term hacker) were also stored as separate concepts.

To search and retrieve terminology data from the database, a web service (called heterogeneity service, see Mayr & Walter, 2008) was built to support cross-concordance searches for individual start terms, mapped terms, start and destination vocabularies as well as different types of relations.

Many cross-concordances are already utilized for search in the German Social Science Information Portal sowiport[59], which offers bibliographical and other information resources (incl. 15 databases with 10 different vocabularies and about 2.5 million bibliographical references). The application, which uses the equivalence relations[60], looks up search terms in the controlled vocabulary term list and then automatically adds all equivalent terms from all available vocabularies to the query. If the controlled vocabularies are in different languages, the heterogeneity service also provides a translation from the original term to the preferred controlled term in the other language. If the original query contains a Boolean command, it remains intact after the query expansion (i.e. each query word gets expanded separately). In the results list, a small icon symbolizes the transformation for the user (see FIG. 2).



Ihre Suche: Überall: [**handlungsforschung** ◀▶ und **management** ◀▶]

◀▶ **Termtransformation**

Der Term: handlungsforschung wurde transformiert in:
Aktionsforschung

◀▶ **Termtransformation**

Der Term: management wurde transformiert in:
Betriebsführung oder
Betriebswirtschaft oder
Betriebswirtschaftslehre oder
Geschäftsführung oder

FIG. 2. Term mapping for search in sowiport. All terms are added to the query with a Boolean OR.

Because of performance issues, the cross-concordance query expansion doesn't distinguish between different databases and their preferred controlled vocabulary terms given a concept, but adds all equivalent terms to the query. In principle, this use of the terminology network expands a query with synonyms or quasi-synonyms of the original query terms. By adding terms to the query, recall should increase, that is, more relevant documents will be found. It is unclear, however, whether the indiscriminate expansion of the original query without regard for the terms' appropriateness for a given database can actually decrease the precision of the search. If the created equivalence mappings only denote correct synonyms, then the adding of true synonyms should have no such effect. However, homonymic terms as well as slight variations in the meaning of a concept can have a detrimental impact on the quality and precision of the query. In an ideal case, the searcher could be represented with a selection of terms garnered from the cross-concordances and then select an appropriate formulation. As most users prefer simple search interfaces with quick results (Jansen & Pooch, 2000; Bandos & Resnick, 2004), an interactive search process or even an appropriate visualization of the cross-concordance work is difficult to accomplish.

Another major issue for a growing terminology network is the scale and overlap of cross-concordances. The more vocabularies are mapped to each other, the more terms occur multiple times in variant mappings[61], which makes automatic query expansion more imprecise. On the other hand, the more vocabularies are added in such a network, the more inferences can be drawn for additional mappings. Indirect mappings via a pivot vocabulary could help in connecting

[59] http://www.sowiport.de/

[60] The other relations, which can lead to imprecise query formulations because they are broader, narrower or related to the original term, could be leveraged in an interactive search, when the searcher can guide and direct the selection of search term.

[61] For example: term A from vocabulary 1 also occurs in vocabulary 2. A variant mapping exists when term A from vocabulary 1 is mapped to term B in vocabulary 3, but term A from vocabulary 2 is mapped to term C in vocabulary 3. This might be the correct mapping because the concepts in the different vocabularies are differently connotated but most of the time this will introduce noise to the network.

vocabularies that haven’t been mapped to each other. A sufficiently large network could assist in reducing the mapping errors introduced by statistical or indirect mappings.

4. Leveraging a Terminology Network – Outlook

This project is the largest terminology mapping effort in Germany. The number and variety of controlled vocabularies targeted provide an optimal basis for further research opportunities. To our knowledge, terminology mapping efforts and the resulting terminology networks have rarely been evaluated with stringent qualitative and quantitative measures.

The current cross-concordances will be further analyzed and leveraged for distributed search not only in the sowiport portal but also in the German interdisciplinary science portal vascoda. The terminology mapping data is made available for research purposes. Some mappings are already in use for the domain-specific track at the CLEF (Cross-Language Evaluation Forum) retrieval conference (Petras, Baerisch & Stempfhuber, 2007).

We also plan on leveraging the mappings for vocabulary help in the initial query formulation process as well as for the ranking of retrieval results (Mayr, Mutschke & Petras, 2008).

Aside from its application in a distributed search scenario, the semantic web community might be able to find new and interesting usages for terminology data like this one. The SKOS standard (Simple Knowledge Organization System)[62] contains a section on mapping vocabularies in its draft version. Once the standard gets stabilized, we plan on transferring the cross-concordance data to the SKOS format. If more vocabularies and mappings become available in SKOS, then further research into connecting previously unmapped terminology networks with each other should be possible.

Acknowledgements

The project was funded by BMBF, grant no. 01C5953. We wish to thank all our project partners for their collaboration. We appreciate the help of our colleagues Anne-Kathrin Walter, who implemented the database and the heterogeneity service prototype, and Stefan Bärisch, who integrated the cross-concordances in the current sowiport search.

References

Bandos, Jennifer A., and Marc L. Resnick. (2002). Understanding query formation in the use of Internet search engines. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Conference,* (pp. 1291-1295).

Hellweg, Heiko, Jürgen Krause, Thomas Mandl, Jutta Marx, Matthias N. O. Müller, Peter Mutschke, et al. (2001). *Treatment of semantic heterogeneity in information retrieval.* Bonn: IZ Sozialwissenschaften.

IFLA (2005). *Guidelines for multilingual thesauri*, Working Group on Guidelines for Multilingual Thesauri. Classification and Indexing Section, IFLA.

ISO 5964 (1985). *Documentation - Guidelines for the establishment and development of multilingual thesauri.*

Jansen, Bernhard J., and Udo Pooch. (2000). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology,* 52(3), 235-246.

Krause, Jürgen. (2003). Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. *Paper presented at the IFLA 2003, World Library and Information Congress*: *69th IFLA General Conference and Council, Berlin.*

Mayr, Philipp, Peter Mutschke, and Vivien Petras. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. *Library Review, 57*(3), 213-224.

Mayr, Philipp, and Vivien Petras. (2008 to be published). Cross-concordances: Terminology mapping and its effectiveness for information retrieval. *Paper to be presented at the74th IFLA World Library and Information Congress Québec, Canada*. Retrieved from http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf.

[62] http://www.w3.org/2004/02/skos/

Mayr, Philipp, and Anne-Kathrin Walter. (2008). Mapping knowledge organization systems. In H. Peter Ohly, Sebastian Netscher, and Konstantin Mitgutsch (Eds.), *Fortschritte der Wissensorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation,* (pp. 80-95). Würzburg: Ergon.

Patel, Manjula, Traugott Koch, Martin Doerr, and Chrisa Tsinaraki. (2005). *Semantic Interoperability in Digital Library Systems*.

Petras, Vivien, Stefan Baerisch, and Maximillian Stempfhuber. (2007). The Domain-Specific Track at CLEF 2007. *Cross Language Evaluation Forum Workshop (CLEF) 2007, Budapest.*

Zeng, Marcia Lei, and Lois Mai Chan. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, *55*(3), 377-395.

Cool URIs for the DDC: Towards Web-Scale Accessibility of a Large Classification System

Michael Panzer
OCLC, USA
panzerm@oclc.org

Abstract

The report discusses metadata strategies employed and problems encountered during the first step of transforming the DDC into a Web information resource. It focuses on the process of URI design, with regard to W3C recommendations and Semantic Web paradigms. Special emphasis is placed on usefulness of the URIs for RESTful web services.

Keywords: Dewey Decimal Classification; metadata; Uniform Resource Identifiers; web service architecture; classification systems; World Wide Web; REST

1. Introduction

The Dewey Decimal Classification (DDC)[63] system, if it wants to stay relevant to its present and to embrace future users, will have to face the challenge to build a presence on the (Semantic) Web that is not only actionable, but also convenient and useful to its participants. Existing on the Web is the first and currently most important step to potentially become part of “higher-level Web artifacts” that are being built “out of existing Web parts” (T. V. Raman).

Some advances in putting *bibliographic* data and standards on the Web are indeed visible. WorldCat identifiers (OCLC numbers minted as URIs in the worldcat.org namespace) are forming the basis of globally scoped manifestation identifiers for library material; the Library of Congress has recently added permalinks to its catalog records. With regard to subject authority metadata, however, most initiatives keep a very low profile, despite the fact that terminologies, controlled vocabularies, taxonomies, etc., are among the most valuable (and costly) assets of the library community. The relevance of controlled vocabularies for bibliographic standards has become the focus of recent discussion (Coyle & Hillmann, 2007).

The tools and formats that allow those knowledge organization systems to become part of the Semantic Web are emerging. It is now up to providers to rethink historically grown knowledge organization systems (KOS) in terms of these new technologies and make them available for recombination and reuse.

2. Paradigms of Identification, Location, Access

For a resource to be visible on the Web, the single most important piece of information is its URI. It weaves Web resources into the Semantic Web; it connects “things” with information resources describing them, binds information resources together, and (via http) provides information about their relationships. In short, it provides “scaffolding” as well as acts as a “micro-billboard” (Stuart Weibel) for resources.

A URI (Berners-Lee, Fielding, & Masinter, 2005) is commonly defined as a string of characters used to identify or name a single resource. This definition seems odd given the fact that the architecture of the World Wide Web is mainly concerned with representations of *information* resources. Yet, as it is very useful to assign URIs to things that may not be information resources, the discussion about whether the re-entry of the distinction “information

[63] DDC, Dewey, Dewey Decimal Classification, WebDewey, and WorldCat are registered trademarks of OCLC Online Computer Library Center, Inc.

resource/non-information resource” should be allowed into the system of the Web leads to what is now known as its “identity crisis”. It was essentially resolved by simply not allowing this re-entry, making the system/environment distinction, but at the same time leaving the environment as an unmarked state. Therefore, the objects identified by URIs are either information resources or things that may or may not be information resources, i.e.,, more plainly, anything.

Among different URI schemes, choosing http is considered best practice for the Semantic Web, because it “can be resolved by any client without requiring the use of additional plug-ins or client setup configuration” (Berrueta & Phipps, 2008, sec. Naming). (Therefore, the resulting URIs are, in fact, URLs.) An information resource returns representations of the identified resource in response to http requests, a process called dereferencing.

Minting URIs for the DDC is not without complications. While other KOS, mostly thesauri, have been using some kind of internal identification for some time that they now might surface, the situation for the DDC is quite the opposite. It was built from the start upon a set of visible identifiers, the Dewey numbers, which should feature prominently in every URI scheme, even if they need to be augmented considerably to satisfy modern standards of Web architecture.

A naming scheme has to be adopted that both exposes the structure of the DDC for addressability and reference and makes sense to agents (clients) using the Web service by asking questions about DDC resources. To put it a different way: The scheme has to be specific to the DDC as well as adhere to the expectations (i.e.,, standards) of the general and the Semantic Web.

The initial questions are: What taxonomy-level and concept-level metadata elements provided by the DDC should be included in the URI (Mendelsohn & Williams, 2007)? How easy should it be to construct an identifier based on previous classification data, e.g., tag 082 in MARC Bibliographic records? How semantically loaded should they be?

The Web community has quite different approaches when it comes to URI design. Tim Berners-Lee, for example, in his “Axiom of URI opacity”, states that URIs must not contain any elements that can be connected to the resource in a meaningful way, as such elements might raise expectations about the representation that may or may not be fulfilled upon dereferencing the URI. Since URIs are often implemented as late-binding, (practically) nothing about the information resource referenced by the URI should be inferred until the identifier is dereferenced and its representation is retrieved (W3C Technical Architecture Group, 2004, sec. 2.5).

This axiom or – rather – best-practice recommendation is meant to discourage the derivation of metadata from general data of unknown status (“sniffing”). Metadata that can be acquired this way is often closely related to the document or representation of the resource rather than the resource itself. In addition, data elements in URIs are categorized as “external reference metadata”, which is deemed to be the least authoritative metadata source in the context of Web architecture (Fielding & Jacobs, 2006). This type of metadata might depend on not only the intrinsic characteristics of the resource, but also technicalities, media types, publication cycles, etc.

This observation seems to be especially relevant to the DDC, as its metadata will be undergoing significant changes in the near future, the switch to MARC as representation format only being the most obvious. A more subtle change is the way the concept of “editions” is reassessed to signify time-stamped snapshots of the Dewey database without wholesale changes to the referenced resources, rather than adhering to the 7-year cycle of the print edition. This conceptual change is significant to facilitate contiguous ranges of historic versions for individual concepts that can be identified and exposed for retrieval systems (Tennis, 2006).

A second (more moderate) position mandates to include only “well-behaved” metadata that is functionally dependent on the Web document, for example, is unlikely to change independently of the identified resource. In case such metadata changes, it would automatically describe a new document that in turn justifies a new URI.

On the other end of the spectrum are axioms put forward by Roy Fielding's REST (Representational State Transfer) paradigm. He states in his seminal work that it must at least be possible to *treat* URIs as opaque or mere identifiers when dereferenced. Yet the URI is most importantly a resource identifier, not a document identifier.

> [A]uthors need an identifier that closely matches the semantics they intend by a hypermedia reference, allowing the reference to remain static even though the result of accessing that reference may change over time. [… REST is] defining a resource to be the semantics of what the author intends to identify, rather than the value corresponding to those semantics at the time the reference is created. (Fielding, 2000)

This slight redefinition fits into the REST framework that aims at using URIs to actively expose and manipulate resources and their states.

While Berners-Lee emphasizes the character of the URI as a rigid and arbitrary designator, the second position concentrates on it being a locator of documents on a network, and only the third position frames the URI as a concept that allows its representations to be accessed and manipulated in various ways. In addition, RESTful URIs are considered representation-agnostic, so the way in which the data is presented will not interfere with the semantics that govern the identification of a resource.

3. URIs for the Dewey Decimal Classification

When Andy Houghton and colleagues from OCLC's Office of Research started designing a URI structure for the DDC, the result was a very elegant URI Template:

`http://dewey.info/{aspect}/{object}/{locale}/{type}/{version}/{resource}`[64]

Examples of identifiers generated by this template include `http://dewey.info/concept/338.4/en/edn/22` that retrieves or identifies the 338.4 concept in the English version of edition 22. These URIs have some very distinct advantages in being clearly structured, hackable, and (almost) entirely derivable from existing metadata, among others. They also had some drawbacks, however, in being very closely tied to a specific entity-relationship representation of DDC's conceptual structure, and based on an early draft of the URI Template specification that didn't allow for much flexibility in specifying optional and mandatory elements; e.g., segments in the path could not be skipped, only successively omitted starting from the last element. Removing an element in that manner widens the information context of the identifier (determined by the data model that was used to establish the sequence).

From a services perspective, however, this approach seems not flexible enough in the way it mandates what pieces of information agents have to possess in order to interact with the exposed resource. The identifier does not need to be an exact mapping of the data structure of the whole classification; it rather should encourage multiple views on a resource.

The feedback we have received based on the original proposal suggests that the Dewey number, even if semantically not unproblematic, should be the central part of the URI structure. Furthermore, assuming that Dewey concepts, identified by their class number, ought to have the same intension across translations, locale or language could be removed from the concept identifier altogether and handled like any other representation variant. Thirdly, thinking from a

[64] The value set of the {aspect} associated with an {object} contains at least "concept", "scheme", and "index"; {object} is a type of {aspect}, {locale} identifies a Dewey translation, {type} identifies a Dewey edition type and contains, at a minimum, the values "edn" or "abr", {version} identifies a Dewey edition version, {resource} identifies a resource associated with an {object} in the context of {locale}, {type}, and {version}.

REST perspective, identifying resources is closely interrelated to the conception of a service architecture that answers an agent’s questions about those resources.

It should not be a prerequisite to already have a clear conception about the versioning conventions of DDC concepts. If we redefine editions as being nothing more than named time slices, opaque version labels assigned to a group of resources at a specific point in time, the hierarchical `{edition_type}/{edition_version}`, e.g., `edn/22`, `abr/14`, should be represented together in a more generic way as {edition_stamp} with a larger value set (“e22” for the full edition 22, “a14” for the abridged edition 14, “qr-3-2007” for the third quarterly release in 2007 of the Dewey database, or “[2007, 05, 25]” for a specific point in time that would be mapped to the most appropriate version by the service).[65]

Evolving the “edition_type” aspect to a timestamp aspect is useful on another level. With “edition_stamp” becoming just a different moniker for “time”, it can be handled as yet another representational variant of a resource, alongside the representation format specified by HTTP Content-Type.

Following a similar strategy, if Dewey classes have stable intensions independently of language instantiation, the language should be handled in a similar fashion as well. Just like format as the third dimension in which a representation can vary (SKOS, MARC, HTML, etc.), the language/locale element becomes either part of the configuration of the service, query string parameter, or content negotiation. (After abstracting out language, format, and time, we arrive at what is often called a “generic resource” [Berners-Lee, 2000], addressed below in more detail.)

Using the latest draft of the URI Template specification (Gregorio, Hadley, Nottingham, & Orchard, 2008), the new structure looks like this:

```
http://dewey.info/{aspect}{-opt|/|aspect}{object}{-opt|/|object}
   {-list|/|edition_stamp}{-opt|/|edition_stamp}/{-list|/|resource}
```

Let’s analyze some concrete URIs generated by expanding this template:

```
http://dewey.info/class/338.4/2007/05/25/about.en.html
http://dewey.info/class/338.4/e22/about.en.html
```

The above URIs both identify or retrieve an English HTML representation of the 338.4 concept found in edition 22.

```
http://dewey.info/class/2--74-79/2007/05/25/about
http://dewey.info/class/2--74-79/about
```

Format and language of the retrieved resources will be determined by the agent, either by content-negotiation, parsing the generic resource for RDF statements indicating available variants, or using a URI of a fixed resource.

Identifiers for other entities are built accordingly by modifying {aspect} and/or {object}[66]:

[65] Depending on the implementation, it could still be necessary to keep a mechanism to distinguish full and abridged versions independently of how their respective editorial state is labeled, for example `{edition_type}/{edition_stamp}` with {edition_type} being either “abridged” or “full”, and {edition_stamp} similar as explicated above.

[66] Besides “class”, which should only address assignable concepts, {aspect} might include at this point “manual”, “index”, “table”, “scheme”, and “id”.

http://dewey.info/table/1/a14/about.en.skos

http://dewey.info/scheme/about

The first URI identifies a fixed representation of table one, the second URI is the generic identifier for the whole scheme, similar to `dcterms:DDC` defined by the DCMI metadata terms.

So far the `{resource}` has always just been `/about`, indicating a description of the concept found in the DDC. Following the REST paradigm, however, we can weave into the URIs collections of resources that are far more useful for services than just retrieving atomic concepts.[67]

http://dewey.info/class/338.4/e22/ancestors/about

http://dewey.info/class/338.4/ancestors/about.en.skos

Both URIs could be used to identity or retrieve the entire graph of the upward hierarchy of the given concept. The first, identifying a generic representation of the resource, could use content-negotiation and redirecting to HTML by default. Depending on service architecture decisions, a HTTP response code 300 (Multiple Choices) might be returned instead with RDF statements enumerating the choices. The second URI, while retrieving the superordinate concepts of all historic versions of the resource in English, includes links to the content in all other available languages (Raman, 2006).

Depending on what is identified as useful resources for a "Classify API", more application scenarios or use cases, like browsing, retrieval, or query expansion could be supported, by using `/children` (retrieving all immediate subclasses), `/siblings` (returning all coordinate classes with the same superclass, effectively providing a shortcut for a BT/NT traversal or subsequent requests for `/parent` and `/children`), `/related/about?degree=x` (providing the graph of referenced terms up to a specific degree. A `/search` resource resulting from e.g. a keyword search of a collection of all concepts in DDC 22 could be expressed in the same manner: `http://dewey.info/scheme/e22/search/about.de?kw=...`

4. Generic Resources

As the described scheme may produce several URIs that describe the same Dewey concept in somewhat different ways, it is desirable to be able to distinguish a canonical URI or representation (in this context sometimes called a "generic resource").

As discussed above, the definition of information resources is crucial to the architecture of the Web. But since anything might be identified by a URI, there has to be a way to indicate that a URI might denote something other than an information resource. As the resolution of the "httpRange-14 problem" the W3C TAG has decided that when dereferencing a URI and its resource can't be represented by a "message", i.e., identifies not an information resource or a "Web document", a HTTP response 303 (See Other) should be issued pointing to a description of the original resource. For specific ways of addressing these issues in practice, see Sauermann & Cyganiak (2008).

The question for our specific case is now: Is a Dewey class (or concept) an information resource? The Dublin Core Metadata Initiative defines it as a "set of conceptual resources", the DDC Glossary as a "group of objects," SKOS defines "conceptual resources" (a shorthand for concepts) as "units of thought."

[67] See for example (Binding & Tudhope, 2004). The authors, after evaluating different APIs for distributed KOS access, criticize the fact that most APIs mimic the data structure of the KOS too closely and don't support advanced operations like "chunking", i.e., the retrieval of a defined set of concepts with one request to the server.

Steering clear of the intricate philosophical problem (dating back to Maxwell's Demon) if a *group* of things constitutes an information resource while the things alone do not, the cited sources all suggest that concepts of a KOS should be treated as abstract objects (not as information resources). To represent that fact, the {resource} segment has been introduced into the URI to distinguish between the abstract DDC concept and a description of that concept. While `http://dewey.info/class/338.4` indentifies the concept, `http://dewey.info/class/338.4/about` identifies the information resource describing this concept. Since this last URI is designed to be representation-agnostic and provides links to more specific resources, it is in fact the generic representation of this resource.

The benefit of pointing the agent to a generic information resource before negotiating the contents of the representation is mainly semantic. By using this technique it is made clear that all descriptions of the identified resource are variants of the same representation and roughly convey the same information. The relationship of each of those resources to the generic resource is such that they specify one or more dimensions of its genericity.

For example, in our context `http://dewey.info/class/338.4/about` exemplifies an identifier for a generic resource, being *about* an abstract concept. The relationship between the URI and the representation of the resource it identifies may change over time, with respect to language and format requested. The use of the same URI will still be valid, however, because these new resources are considered more specific versions of the generic resource, and their respective relationships would be given as RDF statements about the dimension they specify. On the other hand, the resource that `http://dewey.info/class/338.4/2008/04/03/about` identifies or retrieves is only time-invariant but language- and format-generic, whereas for `http://dewey.info/class/338.4/2008/04/03/about.en.skos` it is completely fixed.

It should also be noted in this context that removing the language from the concept URI implies that a specific language version of a DDC concept can never be addressed as an abstract concept, but only as an information resource describing the abstract (language neutral) concept.

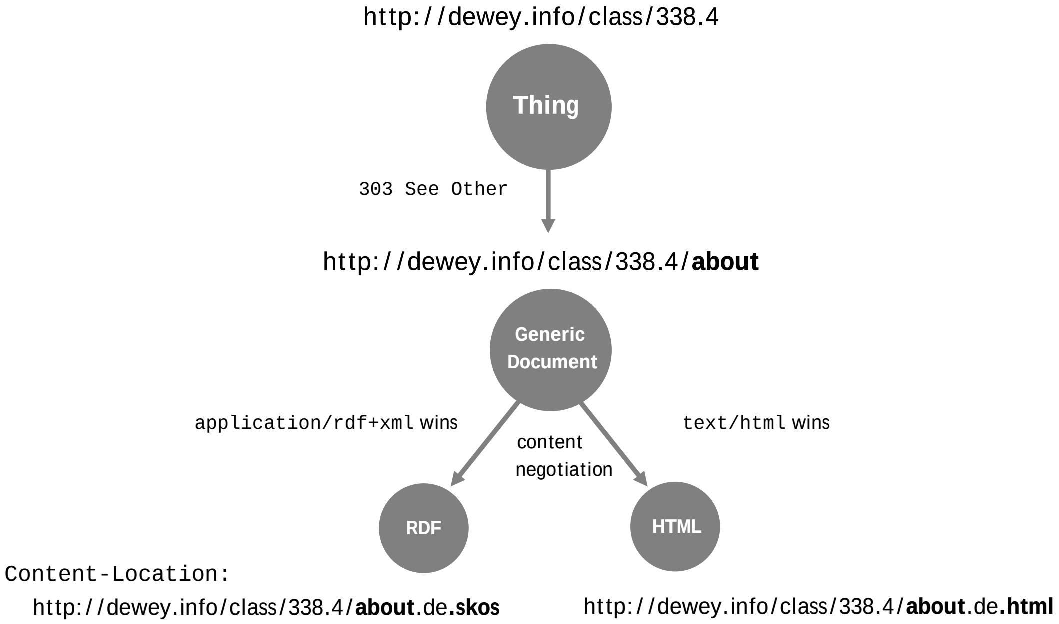

http://dewey.info/class/338.4
Thing
303 See Other
http://dewey.info/class/338.4/**about**
Generic Document
application/rdf+xml wins
content negotiation
text/html wins
RDF
HTML
Content-Location:
http://dewey.info/class/338.4/**about**.de**.skos**
http://dewey.info/class/338.4/**about**.de**.html**

FIG. 1. Generic resources as web documents.

The concept of generic resources is especially important for designating a canonical URI for a given set of resources/representations. The findings above suggest that a candidate for a canonical URI should identify the most generic in a set of resources that can be grouped together as variants of each other.

5. Next steps

There are numerous DDC entities that have not been addressed so far and will therefore not be addressable by the URIs shown above. That doesn't mean that they won't be accessible to applications, however. Even if one assumes that these entities might be irrelevant from a service perspective, it would perhaps be useful to achieve higher granularity for users of the full Dewey data file; and in representation formats like SKOS, every reference has to be a URI, anyway. One possibility would therefore be to use opaque URIs in the `http://dewey.info/id` namespace in parallel, which, for all entities that already have other identifiers, would have to be handled as URI aliases. This set could correspond directly and exhaustively to entities in the Dewey database as represented in MARC Classification and Authorities formats, its entities could be related by OWL and even be used publicly for permalinks.

Another solution: the proposed scheme might be extended by adding fragment identifiers, enabling access to specific pieces of information beyond the level of the suggested URIs, for example, `http://dewey.info/class/1--012/e22/about#caption` to just indentify the caption "Classification" of that class, but these specific entities might be misleading if applied across different data formats (W3C Technical Architecture Group, 2004, sec. 3.2.2), e.g., MARC Classification vs. SKOS. Another potential drawback is that fragment identifiers are stripped from the URI by the user agent, so a service endpoint will never see them.

The usefulness of "shortcuts" has to be addressed in general as well. Every time a "default" is introduced, the expressiveness of the scheme is impoverished by *de facto* defining URI aliases for some resources. If `http://dewey.info/concept/338.4` defaults today (using my current Web browser) to the same representation that is retrieved by `http://dewey.info/concept/338.4/2008/04/04/about.en.html`, the possibility is lost to use the original URI as a canonical identifier for the 338.4 concept independently of time, language, or format. Yet such an identifier is a powerful tool that could retrieve all information about translations, former versions of this concept etc. as OWL or RDF expressions, making it possible for an agent to just work from this resource for any given concept. A better general way of indicating shortcuts would be to interpret an unspecified {aspect} segment as trigger for defaulting behavior, for example: only `http://dewey.info/338.4` would be defined as an alias of the fixed resource shown above, but not `http://dewey.info/concept/338.4.`

References

Berners-Lee, Tim. (2000). Web architecture: Generic resources. Retrieved April 3, 2008, from http://www.w3.org/DesignIssues/Generic.html.

Berners-Lee, Tim, Roy Fielding , and Larry Masinter. (2005). *Uniform Resource Identifier (URI): Generic Syntax*. Standard, IETF. Retrieved from http://www.ietf.org/rfc/rfc3986.txt.

Berrueta, Diego, and Jon Phipps (2008). Best practice recipes for publishing RDF vocabularies. *Working Draft, W3C.* Retrieved from http://www.w3.org/TR/2008/WD-swbp-vocab-pub-20080123/.

Binding, Ceri, and Douglas Tudhope. (2004). KOS at your service: Programmatic access to knowledge organisation systems. *Journal of Digital Information, 4*(4). Retrieved from http://journals.tdl.org/jodi/article/view/jodi-124/109.

Coyle, Karen, and Diane Hillmann. (2007). Resource Description and Access (RDA): Cataloging rules for the 20th century. *D-Lib Magazine, 13*(1/2). Retrieved from http://dlib.org/dlib/january07/coyle/01coyle.html.

Fielding, Roy (2000). *Architectural styles and the design of network-based software architectures.* Thesis (Ph. D., Information and Computer Science), University of California, Irvine. Retrieved from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

Fielding, Roy, and Ian Jacobs. (2006). Authoritative metadata. *TAG Finding, W3C*. Retrieved from http://www.w3.org/2001/tag/doc/mime-respect.html.

Gregorio, Joe, M. Hadley, M. Nottingham, and D. Orchard. (2008). URI Template. *Internet-Draft, IETF*. Retrieved from http://www.ietf.org/internet-drafts/draft-gregorio-uritemplate-03.txt.

Mendelsohn, Noah, and Stuart Williams. (2007). The use of metadata in URIs. *TAG Finding, W3C*. Retrieved from http://www.w3.org/2001/tag/doc/metaDataInURI-31-20070102.html.

Raman, T. V. (2006). On Linking Alternative Representations To Enable Discovery And Publishing. *TAG Finding, W3C.* Retrieved from http://www.w3.org/2001/tag/doc/alternatives-discovery-20061101.html.

Sauermann, Leo, and Richard Cyganiak. (2008). Cool URIs for the semantic web. *Working Draft, W3C*. Retrieved from http://www.w3.org/TR/2008/WD-cooluris-20080321/.

Tennis, Joseph T. (2006). Versioning concept schemes for persistent retrieval. *Bulletin of the American Society for Information Science and Technology, 32*(5), 13–16. doi: 10.1002/bult.2006.1720320506

W3C Technical Architecture Group. (2004). Architecture of the world wide web, volume one. *Recommendation, W3C*. Retrieved from http://www.w3.org/TR/2004/REC-webarch-20041215/.

The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO)

Barbara Levergood
Goettingen State and University Library, Germany
levergood@mail.sub.uni-goettingen.de

Stefan Farrenkopf
Goettingen State and University Library, Germany
farrenkopf@mail.sub.uni-goettingen.de

Elisabeth Frasnelli
Library of the Free University of Bozen-Bolzano, Italy
Elisabeth.Frasnelli@unibz.it

Abstract

The CACAO Project (Cross-language Access to Catalogues and Online Libraries) has been designed to implement natural language processing and cross-language information retrieval techniques to provide cross-language access to information in libraries, a critical issue in the linguistically diverse European Union. This project report addresses two metadata-related challenges for the library community in this context: "false friends" (identical words having different meanings in different languages) and term ambiguity. The possible solutions involve enriching the metadata with attributes specifying language or the source authority file, or associating potential search terms to classes in a classification system. The European Library will evaluate an early implementation of this work in late 2008.

Keywords: Multilingual issues; interoperability; Knowledge Organization Systems (KOS) (e.g., ontologies, taxonomies, and thesauri); normalization and crosswalks

1. Introduction

The European Union (EU) has 23 official languages; many more regional and minority languages are spoken in the 27 member states. A 2006 European Commission/Eurobarometer study revealed that "56% of EU citizens are able to hold a conversation in a language other than their mother tongue", "28% state that they master two languages along with their native language", and "approximately 1 in 10 respondents has sufficient skills to have a conversation in three languages".

In this linguistically diverse and multilingual environment in the EU, there is a tremendous need to provide cross-language access to information (i.e., using one language to find information in another). However, European libraries not only do not share a language, they also have no common subject heading system, classification system, authority files, or bibliographic format. Thus, cross-language access to information in library collections is a complex and difficult problem involving not only natural language analysis and translation, but also the mapping of library subject headings, classifications, and bibliographic formats, presenting problems of both syntactic and semantic interoperability.

The CACAO Project (Cross-language Access to Catalogues and Online Libraries), begun in December 2007, is a 24-month targeted project supported by the eContentplus Programme of the European Commission. It is a consortium of nine partners: Cité des sciences et de l'industrie and Xerox Research Centre Europe from France; the Free University of Bozen-Bolzano, CELI, and Gonetwork from Italy; Kórnik Library from Poland; the National Széchényi Library and the Hungarian Academy of Sciences from Hungary; and Goettingen State and University Library of Germany.

The libraries in the CACAO consortium use a total of at least six different subject heading systems (Library of Congress Subject Headings, Schlagwortnormdatei, Słownik języka haseł przedmiotowych Biblioteki Narodowej [National Library Subject Headings Authority Files], Soggettario per i cataloghi delle biblioteche italiane, 2 local systems) and five different classification systems (Basisklassifikation, Göttinger Online-Klassifikation, Regensburger Verbundklassifikation, 2 local systems). Three of the libraries are multilingual libraries.

CACAO will modify and extend work that has already been implemented at the Library of the Free University of Bozen-Bolzano, a multilingual library having major collections in Italian, German, and English, each with its own subject heading system, as described by Bernardi et al. (2006).

This report reviews two of the important metadata-related challenges that CACAO faces involving the specification of the language of the metadata fields, "false friends" and term ambiguity, and discusses our solutions. We begin with a short description of the CACAO architecture.

2. CACAO Architecture

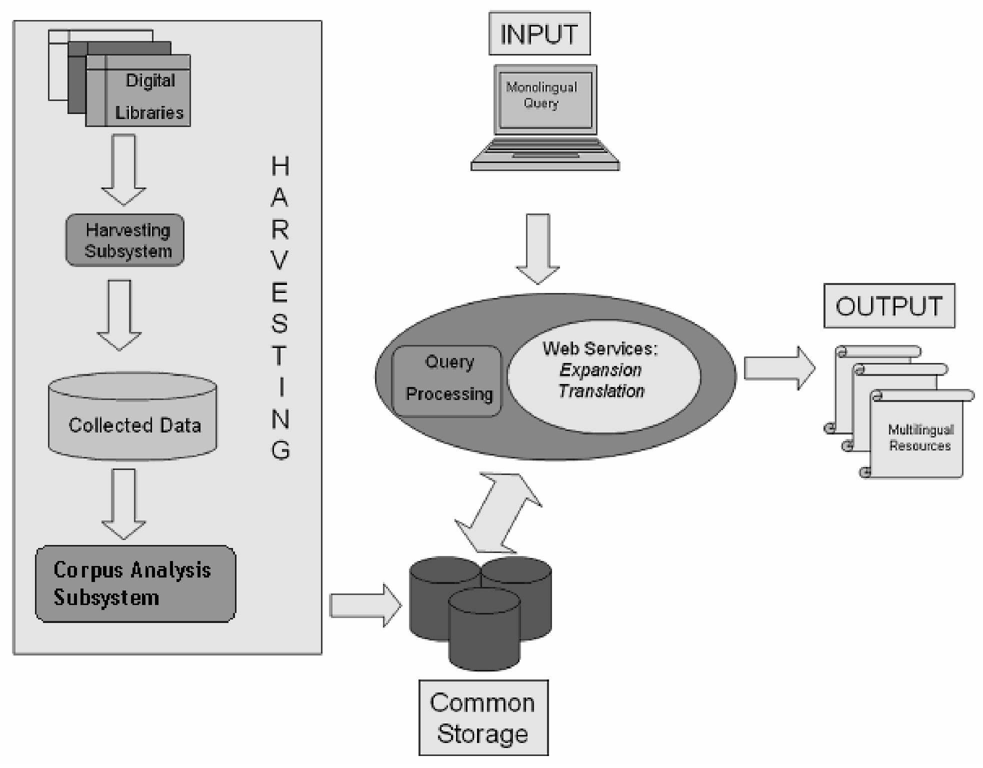


Figure 1 - Architecture Overview (Dini and Bosca (2008), pg. 4)

The CACAO architecture in Figure 1 is designed to support the following vision. A user should be able to enter a monolingual query, say *cat* in English, and retrieve highly relevant records not just in English, but also in any supported language in the database, including records containing, for example, the German word for *cat*, *Katze*, French *chat*, Hungarian *macska*, Italian *gatto*, or Polish *kot*.

As a least-common-denominator solution, CACAO will harvest metadata through library OAI-PMH interfaces, minimally in Dublin Core; MARC 21 may also be accepted if available. The CACAO Corpus Analysis Subsystem performs a variety of analyses on the metadata off-line, the results of which are stored locally and used in support of online Query Processing. When the user enters a query, the Query Processing Subsystem, with the assistance of third-party Web Services providing linguistic analyses, translations, etc., translates and expands the query and matches it against the results of the Corpus Analysis Subsystem. Of course, resources such as lexica, multilingual dictionaries, and thesauri and other controlled vocabularies are accessed by the subsystems.

3. False Friends and Term Ambiguity

We will use a simple example to illustrate some metadata-related problems that arise and some of the possible solutions that we are investigating; these are issues that are challenges not just for CACAO, but also for the library community. Suppose a user enters the query *stove*, wanting to retrieve records containing the English word *stove* or the German translation *Herd*.

USER QUERY: stove

3.1. A Simplistic Solution: Translation

The procedure might seem to be very simple: the Query Processing Subsystem looks up English *stove* in the English-German dictionary, retrieves the German translation *Herd*, and builds a Boolean search query containing those two expressions:

QUERY: stove or Herd

However, this simple query also retrieves false hits containing the English word *herd*:

FALSE HIT: <dc:title>Animal status monitoring and herd management</dc:title>

CORRECT: <dc:title>Herd und Ofen im Mittelalter</dc:title>

CORRECT: <dc:title>The Stove-Top Cook Book</dc:title>

English *herd* and German *Herd* are "false friends", i.e., words in different languages that look similar but that have different meanings. False friends are fairly common, for example English *gift*-German *Gift* ("poison"), English *pain*-French *pain* ("bread"), and English *cane*-Italian *cane* ("dog").

3.2. Solution 2: Enrichment of Metadata

Knowing or being able to determine the language of the terms in a given metadata field increases precision when dealing with false friends. The language would optimally be provided in the metadata itself, as we might find in a German-language catalog which owns the English-language book *The Stove-Top Cook Book* to which German- and English-language subject headings are assigned:

<dc:title xml:lang="en">The Stove-Top Cook Book</dc:title>

<dc:subject xml:lang="de">Herd</dc:subject>

<dc:subject xml:lang="en">Stove</dc:subject>

In the case of a subject term, information about the source of the term in the <dc:subject> field could provide enough information to be able to deduce the language. In this case, we could deduce the language with a fairly high degree of certainty from the fact that the SWD (Schlagwortnormdatei) is a German-language subject heading system:

<dc:subject xsi:type="cacao:SWD">Herd</dc:subject>

This information about the language of the content of the field will be used by CACAO in presenting the ranked results list. Since the German term *Herd* appears in German-language fields in this record:

<dc:title xml:lang="de">Herd und Ofen im Mittelalter</dc:title>

<dc:subject xsi:type="cacao:SWD">Herd</dc:subject>

it would be ranked higher than a record in which the false friend of the German translation of the original search term appears in an English-language field. Alternatively, such a record could be excluded entirely from the results list.

<dc:title xml:lang="en">Animal status monitoring and herd management</dc:title>

3.3. Solution 3: Association to a Class

However, metadata are not always enriched with language or authority attributes as they are in this ideal catalog. CACAO's technical partners are developing a solution for this scenario, the association of terms to a fairly broad class in a library classification system such as the Dewey Decimal Classification (DDC). In our example, the off-line Corpus Analysis Subsystem must have been able to determine that materials about stoves are commonly classed in, e.g., DDC 640 (Home & Family Management), and it has stored this association: stove:DDC 640.

One option would be to organize the results list according to class. For instance, records containing the terms *stove* or *Herd* with a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value provided by the Corpus Analysis Subsystem, 640:

<dc:title>Herd und Ofen im Mittelalter</dc:title>

<dc:subject xsi:type="dcterms:DDC">640</dc:subject>

would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other DDC value for the <dc:subject> element, including records containing the false friend.

<dc:title>Animal status monitoring and herd management</dc:title>

<dc:subject xsi:type="dcterms:DDC">630</dc:subject>

4. Association to a Class and Term Ambiguity

The association to a class technique is used in information retrieval and in CACAO for an even more common problem: term ambiguity. The English word *pipe*, for instance, is ambiguous, meaning either "a long tube", German *Rohr*, or "a device for smoking", German *Pfeife*. For purposes of exposition, assume that on entering *pipe* as a search query, the user is asked which meaning is intended and that the user selects the meaning "a long tube". Using the association to a class technique, the Corpus Analysis Subsystem has determined that relevant materials are often classed in DDC 690 (Building & Construction).

Again, one option would be to organize the results list according to class, similar to the *stove/Herd* example. Records containing the terms *pipe* or *Rohr* and including a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value provided by the Corpus Analysis Subsystem, 690:

<dc:title>Plumbers and pipe fitters library</dc:title>

<dc:subject xsi:type="dcterms:DDC">690</dc:subject>

would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other value for the <dc:subject> element, including records containing the term *pipe* in its unintended meaning:

<dc:title>The pleasures of pipe smoking</dc:title>

<dc:subject xsi:type="dcterms:DDC">390</dc:subject>

Association to a class can also be used to disambiguate an ambiguous target term. For instance, the English search term *dog* translated into Italian is *cane*. However, Italian *cane* has two senses, "dog" and "cock of a weapon", which would be disambiguated in the same way. Records containing the terms *dog* or *cane* and including a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value 630 (Agriculture) would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other value for the <dc:subject> element, including records containing the term *cane* in its unintended meaning.

5. Conclusion

We have argued that the specification of the language of the metadata field, in addition to that of the document itself, is very important so that metadata can be fully exploited for cross-language purposes or in multilingual settings.

If the metadata do not come with or cannot be enriched with the languages of the fields, then CACAO must rely on the association to a class technique, which will be needed in any case. Association to a class was originally designed for and will be used as a solution to the term ambiguity problem; it is similar to synsets used in WordNet and EuroWordNet, which CACAO may also use. The solution involving association to a class may also work as association to a subject heading, although that would require further preparation and testing.

It is important to note that in the association to a class technique, the CACAO Corpus Analysis Subsystem must be able to associate a term such as English *stove* to some class and then the system must be able to match potential hits containing a term such as *Herd* against that same class. In other words, either the systems must contain the same classification system or their classification or subject headings systems must be mappable to the same system. Thus, CACAO's experience with cross-language access so far strongly supports Koch, Neuroth, and Day (2001); NKOS (2001); Chan and Zeng (2002); Harper and Tillett (2007); and many others in the library community who have discussed the importance of the interoperability of subject vocabularies and of classification systems for information retrieval in cross-domain environments. CACAO will rely on already existing mappings such as those provided by the MACS project (Landry (2004, 2006)), which has worked on mappings for RAMEAU (Bibliothèque nationale de France), Library of Congress Subject Headings (British Library), and Schlagwortnormdatei (Deutsche Nationalbibliothek and Bibliothèque nationale suisse).

For optimal performance, even if the metadata of a given collection does not contain the specification for the language of the field as outlined in section 3.2, the Corpus Analysis Subsystem must still have access to such enriched metadata in order to avoid the false friends problem in its off-line analyses. For instance, if the Corpus Analysis Subsystem must determine which class German *Gift* "poison" is most commonly associated with, then it should avoid analyzing fields in which the English *gift* is found. However, we anticipate that the Corpus Analysis Subsystem will have access to a more extensive stored collection of associations between terms and classes than might be available for a given collection.

A prototype of the CACAO information retrieval system was entered in the CLEF 2008 campaign, providing an opportunity to tune and evaluate the system on cross-language library metadata. CACAO's attention will soon turn to related issues involving metadata exchange and interoperability and thereby further explore the characteristics of Dublin Core in its cross-language duties. The European Library, whose Application Profile is Dublin Core-based, will integrate and evaluate CACAO technologies beginning in late 2008. Furthermore, CACAO libraries will be grouped into a single portal and CACAO will additionally create several thematic portals in order to further develop, demonstrate, and promote CACAO technologies.

Acknowledgements

We would especially like to thank Raffaella Bernardi, Jane Greenberg, Tom Baker, and three anonymous reviewers for their careful comments on earlier versions of this report. Of course, all errors are our own. CACAO is funded under the eContentplus programme, ECP 2006 DILI 510035 CACAO Program eContentplus.

References

Bernardi, Raffaella, Diego Calvanese, Luca Dini, Vittorio Di Tomaso, Elisabeth Frasnelli, and Ulrike Kugler et.al. (2006). Multilingual search in libraries. The case-study of the Free University of Bozen-Bolzano. *Proc. 5th International Conference on Language Resources and Evaluation - LREC 2006, Genova*. Retrieved, 3 April, 2008 from http://www.inf.unibz.it/~bernardi/index.php?page=pub.

CACAO: Cross-language Access to Catalogues and Online Libraries. (2007). *Annex 1: Description of Work.* 17 November 2007.

CACAO: Cross-language Access to Catalogues and Online Libraries. (2008). *CACAO Project.* Retrieved, April 10, 2008, from http://www.cacaoproject.eu/.

Chan, Lois Mai, and Marcia Lei Zeng. (2002). Ensuring interoperability among subject vocabularies and knowledge organization schemes: A methodological analysis. *68th IFLA Council and General Conference, 18-24 August 2002, Glasgow.* Retrieved, April 10, 2008, from http://www.ifla.org/IV/ifla68/papers/008-122e.pdf.

Dini, Luca, and Alessio Bosca. (2008). *Definition of programmatic interfaces for accessing data storage in digital libraries, e-catalogues and OPAC.* CACAO Deliverable D.3.1.

European Commission. (2008). *The official EU languages.* Retrieved, April 3, 2008, from http://ec.europa.eu/education/policies/lang/languages/index_en.html.

European Commission. Eurobarometer. (2006). *Europeans and their Languages* (p. 8). Retrieved, April 3, 2008, from http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

Free University of Bozen-Bolzano. (2007) *Multilingual Search.* Retrieved, April 10, 2008, from http://pro.unibz.it/opacdocdigger/index.asp?MLSearch=TRUE.

Harper, Corey A., and Barbara B. Tillett. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & Classification Quarterly 43*(3/4), 47-68.

Jurafsky, Daniel, and James H. Martin. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice Hall.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). *IFLA Satellite Meeting on Classification and Indexing, 14-16 August 2001, Dublin, Ohio, USA.*

Koninklijke Bibliotheek. (2005-2008). *The European Library*. Retrieved, May 27, 2008, from http://www.theeuropeanlibrary.org/.

Landry, Patrice. (2004). Multilingual subject access: The linking approach of MACS. *Cataloging & Classification Quarterly 34*(3/4), 177-191.

Landry, Patrice. (2006). Multilinguisme et langages documentaires: le projet MACS en contexte européen. *Documentation et Bibliothèques 52*(2), 121-129.

Networked Knowledge Organization Systems (NKOS). (2001). Classification crosswalks: Bringing communities together. *The 4th NKOS Workshop at ACM-IEEE Joint Conference on Digital Libraries (JCDL), 28 June 2001, Roanoke, Virginia, USA.* Retrieved, April 10, 2008, from http://nkos.slis.kent.edu/DL01workshop.htm.

OCLC. (2008). *Dewey Services: Dewey Decimal Classification.* Retrieved, April 10, 2008, from http://www.oclc.org/dewey/.

Princeton University. Cognitive Science Laboratory. *WordNet.* Retrieved, May 28, 2008, from http://wordnet.princeton.edu/.

Roux, Claude. (2008). *User Requirements.* CACAO Deliverable D.7.1.

TrebleCLEF Coordination Action. *The Cross-Language Evaluation Forum (CLEF).* Retrieved, May 27, 2008, from http://www.clef-campaign.org/.

University of Amsterdam. *EuroWordNet.* Retrieved, May 28, 2008, from http://www.illc.uva.nl/EuroWordNet/.

Poster Session

Abstracts

Implementation of Rich Metadata Formats and Semantic Tools using DSpace

Imma Subirats
FAO of the United Nations, Italy
Imma.Subirats@fao.org

Areti Ramachandra Durga Prasad
Indian Statistical Institute, India
ard@drtc.isibang.ac.in

Johannes Keizer
FAO of the United Nations, Italy
Johannes.Keizer@fao.org

Andrew Bagdanov
FAO of the United Nations, Italy
Andrew.Bagdanov@fao.org

This poster explores the customization of DSpace to allow the use of the AGRIS Application Profile metadata standard and the AGROVOC thesaurus. The objective is the adaptation of DSpace, through the least invasive code changes either in the form of plug-ins or add-ons, to the specific needs of the Agricultural Sciences and Technology community. Metadata standards such as AGRIS AP, and Knowledge Organization Systems such as the AGROVOC thesaurus, provide mechanisms for sharing information in a standardized manner by recommending the use of common semantics and interoperable syntax (Subirats et al., 2007).

AGRIS AP was created to enhance the description, exchange and subsequent retrieval of agricultural Document-like Information Objects (DLIOs). It is a metadata schema which draws from Metadata standards such as Dublin Core (DC), the Australian Government Locator Service Metadata (AGLS) and the Agricultural Metadata Element Set (AgMES) namespaces. It allows sharing of information across dispersed bibliographic systems (FAO, 2005). AGROVOC[68] is a multilingual structured thesaurus covering agricultural and related domains. Its main role is to standardize the indexing process in order to make searching simpler and more efficient. AGROVOC is developed by FAO (Lauser et al., 2006).

The customization of the DSpace is taking place in several phases. First, the AGRIS AP metadata schema was mapped onto the metadata DSpace model, with several enhancements implemented to support AGRIS AP elements. Next, AGROVOC will be integrated as a controlled vocabulary accessed through a local SKOS or OWL file. Eventually the system will be configurable to access AGROVOC through local files or remotely via webservices. Finally, spell checking and tooltips will be incorporated in the user interface to support metadata editing.

Adapting DSpace to support AGRIS AP and annotation using the semantically-rich AGROVOC thesaurus transform DSpace into a powerful, domain-specific system for annotation and exchange of bibliographic metadata in the agricultural domain.

References

FAO of the United Nations. (2005). *The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology. Guidelines on Best Practices for Information Object Description*. Retrieved March 30, 2008, from ftp://ftp.fao.org/docrep/fao/008/ae909e/ae909e00.pdf.

Lauser, Boris, Margherita Sini, Gauri Salokhe, Johannes Keizer, and Stephen Katz. (2006). Agrovoc Web Services: Improved, real-time access to an agricultural thesaurus. *IAALD Quarterly Bulletin, 2*. Retrieved March 30, 2008, from ftp://ftp.fao.org/docrep/fao/009/ah767e/ah767e00.pdf.

Subirats, Imma, Irene Onyancha, Gauri Salokhe, and Johannes Keizer. (2007). Towards an architecture for open archive networks in Agricultural Sciences and Technology. *Proceedings of the International Conference on Semantic Web & Digital Libraries 2007*. Retrieved March 30, 2008, from ftp://ftp.fao.org/docrep/fao/009/ah766e/ah766e00.pdf.

[68] Agricultural Information Management (AIMS) Website. Retrieved March 30, 2008, from http://www.fao.org/aims/.

SKOS for an Integrated Vocabulary Structure

Marcia L. Zeng
Kent State University,
USA
mzeng@kent.edu

Wei Fan
China Academy of Sciences,
China
fanwei@mail.las.ac.cn

Xia Lin
Drexel University,
USA
xlin@drexel.edu

Keywords: SKOS; Chinese Classified Thesaurus; integrated schemes

In order to transfer the *Chinese Classified Thesaurus* (CCT) into a machine-processable format and provide CCT-based Web services, a pilot study has been conducted in which a variety of selected CCT classes and mapped thesaurus entries are encoded with SKOS. OWL and RDFS are also used to encode the same contents for the purposes of feasibility and cost-benefit comparison.

CCT is a collected effort led by the National Library of China. It is an integration of the national standards *Chinese Library Classification* (CLC) 4th edition and *Chinese Thesaurus* (CT). As a manually created mapping product, CCT provides for each of the classes the corresponding thesaurus terms, and vice versa. The coverage of CCT includes four major clusters: philosophy, social sciences and humanities, natural sciences and technologies, and general works. There are 22 main-classes, 52,992 sub-classes and divisions, 110,837 preferred thesaurus terms, 35,690 entry terms (non-preferred terms), and 59,738 pre-coordinated headings (*Chinese Classified Thesaurus*, 2005)

Major challenges of encoding this large vocabulary comes from its integrated structure. CCT is a result of the combination of two structures (illustrated in Figure 1): a thesaurus that uses ISO-2788 standardized structure and a classification scheme that is basically enumerative, but provides some flexibility for several kinds of synthetic mechanisms

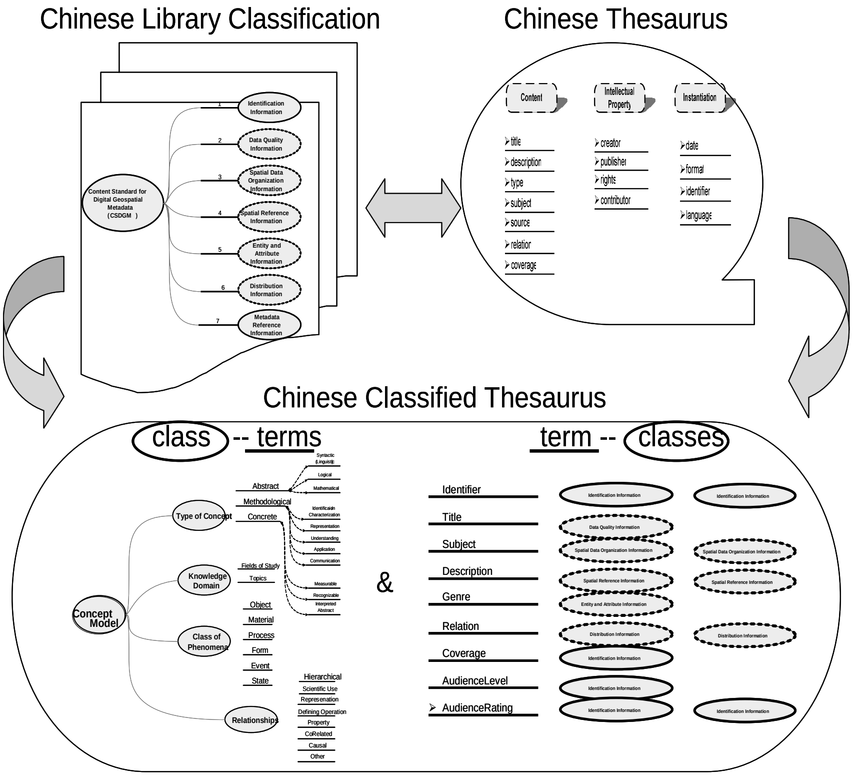


FIG. 1. Illustration of the integrated structure in CCT.

Other challenges include the complex relationships caused by differences of granularities of two original schemes and their presentation with various levels of SKOS elements; as well as the diverse coordination of entries due to the use of auxiliary tables and pre-coordinated headings derived from combining classes, subdivisions, and thesaurus terms, which do not correspond to existing unique identifiers. The poster reports the progress, shares the sample SKOS entries, and summarizes problems identified during the SKOS encoding process. Although OWL Lite and OWL Full provide richer expressiveness, the cost-benefit issues and the final purposes of encoding CCT raise questions of using such approaches.

References:

Chinese Classified Thesaurus. (2005). *National Library of China*. Beijing: Beijing Library Press. Retrieved June 9, 2008, from http://clc.nlc.gov.cn/ztfzfbgk.jsp.

Exploring Evolutionary Biologists' Use and Perceptions of Semantic Metadata for Data Curation

Hollie C. White
University of North Carolina at
Chapel Hill
hcwhite1@email.unc.edu

Keywords: metadata generation; evolutionary biology; personal information management

The wide acceptance of social networking tools in online environments is prompting scientists to engage in metadata creation in not only for organizing their own digital records, but also for contributing to data and journal repositories. Understanding the behaviors and practices of these communities can help us create more effective metadata structures within our information systems.

This point is underscored by information science researchers who have emphasized the need to examine how certain communities interact with, search for, or organize information (Palmer 2001). By examining scientists, information professionals can be more informed in how to create better collections, services, and systems. As library and repository collections become more diverse and personalized, the organization and ingest techniques/applications behind those systems also should be based on observations of how actual user communities work.

One area that is relevant to the practice of scientists and metadata is personal information management (PIM). The study of personal Information management typically focuses on finding (a relative of retrieval), refinding, maintenance, and organization. Metadata is at the core of these activities, although current research seems to focus more on task completion, rather than the underlying metadata structures and arrangements. Most PIM studies and writings have focused on tool development and finding (Jones 2007), but have rarely look closely at the organizational/metadata practices of individuals.

As scientific communities, like evolutionary biology, turn more to cyberinfrastructures for sharing and collaborating with each other, it is important for information professionals to understand the more personal aspects of metadata generation and organization. Recent studies done by the Dryad repository[69] team have looked at different aspects of data sharing and reuse in the evolutionary biology community. These studies have prompted questions about metadata-generation by scientists, their perceptions of the process, and the link between their metadata and the structures imposed in information systems.

This poster will report on a study examining how evolutionary biologists create and use personal metadata to organize their research data. Using an ethnographic interview technique, participants are being interviewed about their current and previous data organization styles and techniques. This information about metadata and information organization can be used to inform new workflow and organization models for knowledge organization and metadata creation practices in developments for repositories, libraries, and cyberinfrastructures.

References

Jones, William. (2007). Personal information management. *Annual Review of Information Science and Technology, 41,* 453-504.

Palmer, Carole. (2001). *Work at the boundaries of science: Information and the interdisciplinary research process.* Drodrecht, Netherlands: Kluwer.

[69] http://datadryad.org/

LCSH is to Thesaurus as Doorbell is to Mammal: Visualizing Structural Problems in the Library of Congress Subject Headings

Simon Spero
UNC Chapel Hill, USA
ses@unc.edu

Keywords: LCSH; Controlled vocabularies; Hierarchical relationships; visualization

The Library of Congress Subject Headings (LCSH) has been developed over the course of more than a century, predating the semantic web by some time. Until the 1986, the only concept-to-concept relationship available was an undifferentiated "See Also" reference, which was used for both associative (RT) and hierarchical (BT/NT) connections. In that year, in preparation for the first release of the headings in machine readable MARC Authorities form, an attempt was made to automatically convert these "See Also" links into the standardized thesaural relations.

Unfortunately, the rule used to determine the type of reference to generate relied on the presence of symmetric links to detect associatively related terms; "See Also" references that were only present in one of the related terms were assumed to be hierarchical. This left the process vulnerable to inconsistent use of references in the pre-conversion data, with a marked bias towards promoting relationships to hierarchical status.

The Library of Congress was aware that the results of the conversion contained many inconsistencies, and intended to validate and correct the results over the course of time. Unfortunately, twenty years later, less than 40% of the converted records have been evaluated.

The converted records, being the earliest encountered during the Library's cataloging activities, represent the most basic concepts within LCSH; errors in the syndetic structure for these records affect far more subordinate concepts than those nearer the periphery. Worse, a policy of patterning new headings after pre-existing ones leads to structural errors arising from the conversion process being replicated in these newer headings, perpetuating and exacerbating the errors.

As the LCSH prepares for its second great conversion, from MARC to SKOS, it is critical to address these structural problems. As part of the work on converting the headings into SKOS, I have experimented with different visualizations of the tangled web of broader terms embedded in LCSH. This poster illustrates several of these renderings, shows how they can help users to judge which relationships might not be correct, and shows just exactly how Doorbells and Mammals are related.

References:

Kiczales, Gregor, Jim Des Rivi`eres, and Daniel Gureasko. (1991). *Bobrow. The art of the metaobject protocol*. Cambridge, Mass.: MIT Press.

Tarjan, Robert. Depth-First Search and Linear Graph Algorithms. (1972) *SIAM Journal on Computing 1*(2), 146-160.

Metadata in an Ecosystem of Presentation Dissemination

R. John Robertson
University of Strathclyde,
United Kingdom
robert.robertson@strath.ac.uk

Phil Barker
Heriot-Watt University,
United Kingdom
philb@icbl.hw.ac.uk

Mahendra Mahey
University of Bath,
United Kingdom
m.mahey@ukoln.ac.uk

Keywords: repository ecology; repository interaction; services; metadata interoperability; management; communication; research dissemination

Developing and managing local practices about metadata implementation (desired quality, workflow, support tools, guidelines, and vocabularies) and about metadata exposure (supported standards, and pre-exposure transformations) requires an ability to understand and communicate the specific complex settings in which the metadata, resources, and users exist. Developing such an understanding is often informed by an implicit or explicit conceptual model.

Ecology is the study of complex natural systems, with the aim of understanding and modeling the processes and interactions between the participants in the system and their environment. The concept is also widely used as a metaphor to describe complex systems within their settings. The Repositories Research Team (which supports repository development work in UK HE) has been examining the use of ecology as a metaphor to support the understanding and representation of interactions between repositories, dependent services, and their users. These interactions whether technical, political, or cultural have a direct impact on the metadata in each repository.

Where many other approaches to modeling facilitate an abstract view of a single type of interaction; the ecologically influenced approach seeks to support communication of the combined influences of a repository's technical and cultural setting, however specific and chaotic (or messy) it may be. The idea that ecology is a suitable metaphor for the interaction of users and technologies has been considered by Davenport (1997), by Nardi and O'Day (2000), in strand of projects funded by the European Union (see Nachira et al., 2007), and by Robertson et al. (2008).

This poster presents an ecologically influenced view of a researcher seeking to disseminate and store their presentations. The interactions and resources that will be considered, as they influence the metadata, include the storage of the presentation in formal and informal services (a repository, SlideShare), different versions of the intellectual content (blog post, slides, paper), different formats (PowerPoint, PDF). Environmental factors, which affect the metadata, that will be considered include influences on the researcher (e.g. availability of web 2.0 tools, the link between career progression and publication of research, a commitment to sharing resources, and institutional policies) and influences on the institutional policies (such as IPR concerns about the use of third party material or the loss of university ownership of intellectual outputs or branding).

References

Davenport, Thomas H. (1997). *Information ecology: Mastering the information and knowledge environment.* Oxford: Oxford University Press.

Nachira, Francesco, Andrea Nicolai, Paolo Dini, Marion Le Louarn, and Lorena Rivera Leon (Eds.). (2007). *Digital Business Ecosystems*. Brussels: European Commission. Retrieved April 9, 2008, from http://www.digital-ecosystems.org/book/de-book2007.html.

Nardi, Bonnie A., and Vicki L. O'Day. (2000). *Information ecologies: Using technology with heart.* Cambridge, Massachusetts: The MIT Press.

Robertson, John R., Mahendra Mahey, and Julie Allinson. (2008). *An ecological approach to repository and service interactions.* Retrieved April 9, 2008, from http://www.ukoln.ac.uk/repositories/digirep/images/a/a5/Introductoryecology.pdf.

A Comparison of Social Tagging Designs and User Participation

Caitlin M. Bentley
Concordia University, Canada
caitlin.bentley@gmail.com

Patrick R. Labelle
Concordia University, Canada
Patrick.Labelle@concordia.ca

Keywords: social tagging; social bookmarking; social computing

Social tagging empowers users to categorize content in a personally meaningful way while harnessing their potential to contribute to a collaborative construction of knowledge (Vander Wal, 2007). In addition, social tagging systems offer innovative filtering mechanisms that facilitate resource discovery and browsing (Mathes, 2004). As a result, social tags may support online communication, informal or intended learning as well as the development of online communities.

The purpose of this mixed methods study is to examine how undergraduate students participate in social tagging activities in order to learn about their motivations, behaviours and practices. A better understanding of their knowledge, habits and interactions with such systems will help practitioners and developers identify important factors when designing enhancements.

In the first phase of the study, students enrolled at a Canadian university completed 103 questionnaires. Quantitative results focusing on general familiarity with social tagging, frequently used Web 2.0 sites, and the purpose for engaging in social tagging activities were compiled. Eight questionnaire respondents participated in follow-up semi-structured interviews that further explored tagging practices by situating questionnaire responses within concrete experiences using popular websites such as YouTube, Facebook, Del.icio.us, and Flickr.

Preliminary results of this study echo findings found in the growing literature concerning social tagging from the fields of computer science (Sen et al., 2006) and information science (Golder & Huberman, 2006; Macgregor & McCulloch, 2006). Generally, two classes of social taggers emerge: those who focus on tagging for individual purposes, and those who view tagging as a way to share or communicate meaning to others. Heavy del.icio.us users, for example, were often focused on simply organizing their own content, and seemed to be conscientiously maintaining their own personally relevant categorizations while, in many cases, placing little importance on the tags of others. Conversely, users tagging items primarily to share content preferred to use specific terms to optimize retrieval and discovery by others.

Our findings should inform practitioners of how interaction design can be tailored for different tagging systems applications, and how these findings are positioned within the current debate surrounding social tagging among the resource discovery community. We also hope to direct future research in the field to place a greater importance on exploring the benefits of tagging as a socially-driven endeavour rather than uniquely as a means of managing information.

References

Golder, Scott A., and Bernardo A. Huberman. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Macgregor, George, and Emma McCulloch. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review, 55*(5), 291-300.

Mathes, Adam. (2004). *Folksonomies - cooperative classification and communication through shared metadata*. Retrieved November 3, 2007, from http://adammathes.com/academic/computer-mediated-communication/folksonomies.html

Sen, Shilad, Shyong Lam, Al M. Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, et al. (2006). Tagging, communities, vocabulary, evolution. *CSCW '06, Banff, Alberta, Canada, 4-8 November, 2006,* (pp. 181-190).

Vander Wal, Thomas. (2007). *Folksonomy coinage and definition.* Retrieved November 3, 2007, from http://vanderwal.net/folksonomy.html

The Data Documentation Initiative (DDI)

Joachim Wackerow
GESIS
Mannheim, Germany
joachim.wackerow@gesis.org

Keywords: Data Document Initiative, DDI, metadata, datasets

The Data Documentation Initiative (DDI)[70] is an international effort to establish an XML-based standard for the compilation, presentation, and exchange of documentation for datasets in the social and behavioral sciences. The most recent version 3.0 of the DDI supports a rich and structured set of metadata elements that not only fully informs a potential data analyst about a given dataset but also facilitates computer processing of the data.[71] Moreover, data producers will find that by adopting the DDI standard they can produce better and more complete documentation as a natural step in designing and fielding computer-assisted interviewing.

DDI 3.0 embraces the full life cycle of the data from conception, through development of the data collection instrument, collection and cleaning of data, production of data products, distribution, preservation, and reuse or analysis of the data. DDI 3.0 is designed to facilitate sharing schemes for concepts, questions, coding, and variables within organizations or throughout the social science research community. Comparison through direct inheritance as in the case of comparison-by-design or through the mapping of items like variables or categories allow capture of the harmonization processes used in creating integrated files in an uniform and machine-actionable way. DDI 3.0 is providing the structural support needed to facilitate comparative survey work in a way that was previously unavailable in an open, non-proprietary system.

A specific DDI module allows for the capture and expression of native Dublin Core elements (DCMES), used either as references or as descriptions of a particular set of metadata. This module uses the simple Dublin Core namespace represented as XML Schema following the guidelines for implementing Dublin Core in XML. In DDI, the Dublin Core is not used as the primary citation mechanism – this module is included to support applications which understand the Dublin Core XML, but which do not understand DDI. This module is used wherever citations are permitted within DDI 3.0 (like citations of a study description or of other material).

DDI 3.0 is aligned with other metadata standards as well: with SDMX (time-series data) for exchanging aggregate data, with ISO/IEC 11179 (metadata registry) for building data registries such as question, variable, and concept banks, and with FGDC and ISO 19115 (geographic standards) for supporting GIS users.

DDI 3.0 is described in a conceptual model which is also expressed in the Universal Modeling Language (UML). Modular XML Schemas are derived from the conceptual model. Many elements support computer processing – that is, it will go beyond being "human readable", and move toward the goal of being "machine-actionable". The final release of DDI 3.0 has been published on April 28th 2008. The standard was developed by the DDI Alliance, an international group encompassing data archives and research institutions from several countries in Western Europe and North America.

Earlier versions of DDI provide examples of institutions and applications: the Inter-university Consortium for Political and Social Research (ICPSR) Data Catalog, the Council of European Social Science Data Services (CESSDA) Data Portal, the Dataverse Network, the International Household Survey Network (IHSN), NESSTAR Software for publishing data on the Web and online analysis, and the Microdata Management Toolkit (by the World Bank Data Group for IHSN).

[70] www.ddialliance.org
[71] http://www.ddialliance.org/ddi3/

junii2 and AIRway - an Application Profile for Scholarly Works and Its Application for Link Resolvers

Kunie Horikoshi
Hokkaido University,
Japan
airway@lib.hokudai.ac.jp

Shigeki Sugita
Hokkaido University,
Japan
airway@lib.hokudai.ac.jp

Yuji Nonaka
Hokkaido University,
Japan
airway@lib.hokudai.ac.jp

Satsuki Kamiya
Hokkaido University,
Japan
airway@lib.hokudai.ac.jp

Izumi Sugita
National Institute of
Informatics,
Japan
izumi@nii.ac.jp

Haruo Asoshina
National Institute of
Informatics,
Japan
asoshina@nii.ac.jp

Keywords: application profile; scholarly works; OpenURL

A large number of scholarly works is self-archived at the university's Open Access repositories.

Researchers can search these materials using general web search engines such as Google, as well as with OAI-PMH-based search engines such as OAIster (http://www.oaister.org/). The archives can also be accessed using federated search services such as MetaLib by setting the repositories as a search target. However, it remains difficult for researchers to access materials in these repositories using standard academic databases such as Thomson Reuters' Web of Science.

The National Institute of Informatics (NII) in Japan has developed a DC application profile called junii2 (http://ju.nii.ac.jp/oai/junii2.xsd) for scholarly works. The AIRway Project (Access path to Institutional Resources via link resolvers) has used this profile to develop a new way of connecting university repositories with academic databases via link resolvers.

junii2 is designed as an OpenURL-compliant schema (info:ofi/fmt:xml:xsd:journal), and has now been widely adopted by more than 70 university repositories in Japan. A particular feature is its ability to describe variant self-archived materials with a version description function (specifying whether it is an author's draft or the final published version) and information on the availability of the full text in the repository.

AIRway is an internet server that harvests metadata from university repositories. After harvesting metadata, AIRway separates the metadata of materials whose full texts are available in the repositories from others. A link resolver sends an OpenURL request to the AIRway server before creating its navigation window. If metadata of the requested material are found in the AIRway server and the material's full text is available in a repository, the AIRway server provides the xml for the metadata of the material to the link resolver. Rather than being a new service system for end users, it is a back-end knowledgebase for existing link resolvers. 1CATE (OCLC's link resolver) and some installations of SFX (Ex Libris' link resolver) now use AIRway as one of their knowledgebases.

In this way, junii2 and AIRway make Open Access scholarly works in university repositories accessible through general academic databases. This will be particularly effective if, for example, someone without a license to access an electronic journal finds a research paper on the journal in the search results of an academic database.

The AIRway Project is funded by the NII Institutional Repositories Program (http://www.nii.ac.jp/irp/en/).

Contact

AIRway Project <airway@lib.hokudai.ac.jp>

Open Identification and Linking of the Four Ws

Ryan Shaw
University of California, Berkeley, USA
ryanshaw@ischool.berkeley.edu

Michael Buckland
University of California, Berkeley, USA
buckland@ischool.berkeley.edu

Platforms for social computing connect users via shared references to people with whom they have relationships, events attended, places lived in or traveled to, and topics such as favorite books or movies. Since free text is insufficient for expressing such references precisely and unambiguously, many social computing platforms coin identifiers for topics, places, events, and people and provide interfaces for finding and selecting these identifiers from controlled lists. Using these interfaces, users collaboratively construct a web of links among entities.

This model needn't be limited to social networking sites. Understanding an item in a digital library or museum requires context: information about the topics, places, events, and people to which the item is related. Students, journalists and investigators traditionally discover this kind of context by asking “the four Ws”: what, where, when and who. The DCMI Kernel Metadata Community has recognized the four Ws as fundamental elements of descriptions (Kunze & Turner, 2007). Making better use of metadata to answer these questions via links to appropriate contextual resources has been our focus in a series of research projects over the past few years. Currently we are building a system for enabling readers of any text to relate any topic, place, event or person mentioned in the text to the best explanatory resources available. This system is being developed with two different corpora: a diverse variety of biographical texts characterized by very rich and dense mentions of people, events, places and activities, and a large collection of newly-scanned books, journals and manuscripts relating to Irish culture and history.

Like a social computing platform, our system consists of tools for referring to topics, places, events or people, disambiguating these references by linking them to unique identifiers, and using the disambiguated references to provide useful information in context and to link to related resources. Yet current social computing platforms, while usually amenable to importing and exporting data, tend to mint proprietary identifiers and expect links to be traversed using their own interfaces. We take a different approach, using identifiers from both established and emerging naming authorities, representing relationships using standardized metadata vocabularies, and publishing those representations using standard protocols so that links can be stored and traversed anywhere. Central to our strategy is to move from appearances in a text to naming authorities to the the construction of links for searching or querying trusted resources.

Using identifiers from naming authorities, rather than literal values (as in the DCMI Kernel) or keys from a proprietary database, makes it more likely that links constructed using our system will continue to be useful in the future. WorldCat Identities URIs (http://worldcat.org/identities/) linked to Library of Congress and Deutsche Nationalbibliothek authority files for persons and organizations and Geonames (http://geonames.org/) URIs for places are stable identifiers attached to a wealth of useful metadata. Yet no naming authority can be totally comprehensive, so our system can be extended to use new sources of identifiers as needed. For example, we are experimenting with using Freebase (http://freebase.com/) URIs to identify historical events, for which no established naming authority currently exists.

Stable identifiers (URIs), standardized hyperlinked data formats (XML), and uniform publishing protocols (HTTP) are key ingredients of the web's open architecture. Our system provides an example of how this open architecture can be exploited to build flexible and useful tools for connecting resources via shared references to topics, places, events, and people.

References

Kunze, John A. and Adrian Turner. (2007). *Kernel Metadata and Electronic Resource Citations (ERCs)*. Retrieved June 13, 2008, from http://www.cdlib.org/inside/diglib/ark/ercspec.html.

Web 2.0 Semantic Systems: Collaborative Learning in Science

Michael Shoffner
Renaissance Computing Institute
UNC Chapel Hill
shoffner@renci.org

Jane Greenberg
Metadata Research Center,
School of Information and Library Science,
UNC Chapel Hill
janeg@email.unc.edu

Jacob Kramer-Duffield
Metadata Research Center,
School of Information and Library Science,
UNC Chapel Hill
jkd@email.unc.edu

David Woodbury
Metadata Research Center,
School of Information and Library Science,
UNC Chapel Hill
dnw@email.unc.edu

Keywords: metadata, shared semantics, Web 2.0, collaborative learning

The basic goal of education within a discipline is to transform a novice into an expert. This entails moving the novice toward the "semantic space" that the expert inhabits—the space of concepts, meanings, vocabularies, and other intellectual constructs that comprise the discipline.

Metadata is significant to this goal in digitally mediated education environments. Encoding the experts' semantic space not only enables the sharing of semantics among discipline scientists, but also creates an environment that bridges the semantic gap between the common vocabulary of the novice and the granular descriptive language of the seasoned scientist (Greenberg, et al, 2005). Developments underlying the Semantic Web, where vocabularies are formalized in the Web Ontology Language (OWL), and Web 2.0 approaches of user-generated folksonomies provide an infrastructure for linking vocabulary systems and promoting group learning via metadata literacy.

Group learning is a pedagogical approach to teaching that harnesses the phenomenon of "collective intelligence" to increase learning by means of collaboration. Learning a new semantic system can be daunting for a novice, and yet it is integral to advance one's knowledge in a discipline and retain interest. These ideas are key to the "BOT 2.0: Botany through Web 2.0, the Memex and Social Learning" project (Bot 2.0).[72]

Bot 2.0 is a collaboration involving the North Carolina Botanical Garden, the UNC SILS Metadata Research center, and the Renaissance Computing Institute (RENCI). Bot 2.0 presents a curriculum utilizing a *memex* as a way for students to link and share digital information, working asynchronously in an environment beyond the traditional classroom. Our conception of a memex is not a centralized black box but rather a flexible, distributed framework that uses the most salient and easiest-to-use collaborative platforms (e.g., Facebook, Flickr, wiki and blog technology) for personal information management. By meeting students "where they live" digitally, we hope to attract students to the study of botanical science. A key aspect is to teach students scientific terminology and about the value of metadata, an inherent function in several of the technologies and in the instructional approach we are utilizing.

This poster will report on a study examining the value of both folksonomies and taxonomies for post-secondary college students learning plant identification. Our data is drawn from a curriculum involving a virtual independent learning portion and a "BotCamp" weekend at UNC, where students work with digital plant specimens that they have captured. Results provide some

[72] http://ils.unc.edu/mrc/bot-20/

insight into the importance of collaboration and shared vocabulary for gaining confidence and for student progression from novice to expert in botany.

Acknowledgments

This work is supported by the National Science Foundation, grant number #0737466.

References

Greenberg, Jane, Brian Heidorn, Stephen Seiberling, and Alan S. Weakly. (2005). Growing vocabularies for plant identification and scientific learning. *International Conference on Dublin Core and Metadata Applications, Madrid, Spain, September 12-15, 2005.*

Doing the LibraryThing™ in an Academic Library Catalog

Christine DeZelar-Tiedman
University of Minnesota
Libraries, USA
dezel002@tc.umn.edu

Keywords: LibraryThing; social tagging; controlled vocabulary; library catalogs

Many libraries and other cultural institutions are incorporating Web 2.0 features and enhanced metadata into their catalogs (Trant 2006). These value-added elements include those typically found in commercial and social networking sites, such as book jacket images, reviews, and user-generated tags. One such site that libraries are exploring as a model is LibraryThing (www.librarything.com) LibraryThing is a social networking site that allows users to "catalog" their own book collections. Members can add tags and reviews to records for books, as well as engage in online discussions. In addition to its service for individuals, LibraryThing offers a fee-based service to libraries, where institutions can add LibraryThing tags, recommendations, and other features to their online catalog records.

This poster will present data analyzing the quality and quantity of the metadata that a large academic library would expect to gain if utilizing such a service, focusing on the overlap between titles found in the library's catalog and in LibraryThing's database, and on a comparison between the controlled subject headings in the former and the user-generated tags in the latter. During February through April 2008, a random sample of 383 titles from the University of Minnesota Libraries catalog was searched in LibraryThing. Eighty works, or 21 percent of the sample, had corresponding records available in LibraryThing.

Golder and Huberman (2006) outline the advantages and disadvantages of using controlled vocabulary for subject access to information resources versus the growing trend of tags supplied by users or by content creators. Using the 80 matched records from the sample, comparisons were made between the user-supplied tags in LibraryThing (social tags) and the subject headings in the library catalog records (controlled vocabulary system). In the library records, terms from all 6XX MARC fields were used. To make a more meaningful comparison, controlled subject terms were broken down into facets according to their headings and subheadings, and each unique facet counted separately. A total of 227 subject terms were applied to the 80 catalog records, an average of 2.84 per record. In LibraryThing, 698 tags were applied to the same 80 titles, an average of 8.73 per title. The poster will further explore the relationships between the terms applied in each source, and identify where overlaps and complementary levels of access occur.

References

Golder, Scott A., and Bernardo A. Huberman. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Trant, Jennifer. (2006). Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. *New Review of Hypermedia & Multimedia, 12*(1), 83-105.

Applying DC to Institutional Data Repositories

Robin Rice
EDINA and Data Library
University of Edinburgh, UK
R.Rice@ed.ac.uk

Keywords: data curation; datasets; DDI; DSpace; institutional repositories; Dublin Core

DISC-UK DataShare (2007-2009)[73], a project led by the University of Edinburgh and funded by JISC (Joint Information Systems Committee, UK), arises from an existing consortium of academic data support professionals working in the domain of social science datasets (Data Information Specialists Committee-UK). We are working together across four universities with colleagues engaged in managing open access repositories for e-prints. Our project supports ‘early adopter’ academics who wish to openly share datasets and presents a model for depositing ‘orphaned datasets’ that are not being deposited in subject-domain data archives/centres.

Outputs from the project are intended to help to demystify data as complex objects in repositories, and assist other institutional repository managers in overcoming barriers to incorporating research data. By building on lessons learned from recent JISC-funded data repository projects such as SToRe[74] and GRADE[75] the project will help realize the vision of the Digital Repositories Roadmap, e.g. the milestone under Data, “Institutions need to invest in research data repositories” (Heery and Powell, 2006).

Application of appropriate metadata is an important area of development for the project. Datasets are not different from other digital materials in that they need to be described, not just for discovery but also for preservation and re-use. The GRADE project found that for geo-spatial datasets, Dublin Core metadata (with geo-spatial enhancements such as a bounding box for the ‘coverage’ property) was sufficient for discovery within a DSpace repository, though more in-depth metadata or documentation was required for re-use after downloading. The project partners are examining other metadata schemas such as the Data Documentation Initiative (DDI) versions 2 and 3, used primarily by social science data archives (Martinez, 2008). Crosswalks from the DDI to qualified Dublin Core are important for describing research datasets at the study level (as opposed to the variable level which is largely out of scope for this project).

DataShare is benefiting from work of of the DRIADE project (application profile development for evolutionary biology) (Carrier, et al, 2007), eBank UK[76] (developed an application profile for crystallography data) and GAP[77] (Geospatial Application Profile, in progress) in defining interoperable Dublin Core qualified metadata elements and their application to datasets for each partner repository. The solution devised at Edinburgh for DSpace will be covered in the poster.

References

Carrier, Sarah, Jed Dube, and Jane Greenberg. (2007). The DRIADE project: Phased application profile development in support of open science. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007,* (pp. 35-42).

Heery, Rachel, and Powell, Andy. (2006). *Digital repositories roadmap: Looking forward*. Retrieved from http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/#roadmap-200604.

Martinez, Luis. (2008). *The Data Documentation Initiative (DDI) and institutional repositories.* Retrieved from http://www.disc-uk.org/docs/DDI_and_IRs.pdf.

[73] http://www.disc-uk.org/datashare.html
[74] http://www.era.lib.ed.ac.uk/handle/1842/1412
[75] http://edina.ac.uk/projects/grade/
[76] http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/
[77] http://edina.ac.uk/projects/GAP_summary.html

Author Index

Subject Index

Metadata is a key aspect of our evolving infrastructure for information management, social computing, and scientific collaboration.

DC-2008 will focus on metadata challenges, solutions, and innovation in initiatives and activities underlying semantic and social applications. Metadata is part of the fabric of social computing, which includes the use of wikis, blogs, and tagging for collaboration and participation. Metadata also underlies the development of semantic applications, and the Semantic Web — the representation and integration of multimedia knowledge structures on the basis of semantic models. These two trends flow together in applications such as Wikipedia, where authors collectively create structured information that can be extracted and used to enhance access to and use of information sources.

Recent discussion has focused on how existing bibliographic standards can be expressed as Semantic Web vocabularies to facilitate the ingration of library and cultural heritage data with other types of data. Harnessing the efforts of content providers and end-users to link, tag, edit, and describe their information in interoperable ways (”participatory metadata”) is a key step towards providing knowledge environments that are scalable, self-correcting, and evolvable.

DC-2008 will explore conceptual and practical issues in the development and deployment of semantic and social applications to meet the needs of specific communities of practice.

DC-2008 Berlin is organised by

KIM Kompetenzzentrum Interoperable Metadaten

Max Planck digital library

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Dublin Core Metadata Initiative®
Making it easier to find information.

DEUTSCHE NATIONAL BIBLIOTHEK

funded by

Deutsche Forschungsgemeinschaft DFG

Bundesministerium für Bildung und Forschung

supported by

WIKIMEDIA DEUTSCHLAND
Gesellschaft zur Förderung Freien Wissens e.V.

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISBN: 978-3-940344-49-6
ISSN: 1939-1358

Universitätsverlag Göttingen