DC-2018
Porto, Portugal

Open Metadata for Open Knowledge

2018 Proceedings of the

International Conference on Dublin Core and Metadata Applications

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

10-13 September, 2018

ISSN: 1939-1366 (Online)

WORKSHOPS

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

CONFERENCES

DC-2001, Tokyo, Japan: 22-26 October 2001
DC-2002, Florence, Italy: 14-17 October 2002
DC-2003, Seattle, Washington, U.S.A.: 28 September - 2 October 2003
DC-2004, Shanghai, China: 10-14 October 2004
DC-2005, Leganés (Madrid), Spain: 12-15 September 2005
DC-2006, Manzanillo, Colima, Mexico: 3-6 October 2006
DC-2007, Singapore: 27-31 August 2007
DC-2008, Berlin, Germany: 22-26 September 2008
DC-2009, Seoul, Korea: 12-16 October 2009
DC-2010, Pittsburgh, Pennsylvania, USA: 20-22 October 2010
DC-2011, The Hague, The Netherlands: 21-23 September 2011
DC-2012, Kuching, Sarawak, Malaysia: 3-7 September 2012
DC-2013, Lisbon, Portugal: 2-6 September 2013
DC-2014, Austin, Texas, USA, 8-11 October 2014
DC-2015, São Paulo, Brazil: 1-4 September 2015
DC-2016, Copenhagen, Denmark: 13-16 October 2016
DC-2017, Washington, DC, USA: 26-29 October 2017
DC-2018, Porto, Portugal: 10-13 September 2018

DC-2018
Welcome

Welcome to DCMI 2018, in Porto, Portugal!

As chair of the DCMI Governing Board, I have the special honour of writing this note to the DCMI community and to welcome you to my country and to the beautiful city of Porto. I hope you enjoy the conference and this city which, after a while, makes such a deep impression on visitors that they are reluctant to leave!

In 2018 we are pleased to host our annual conference in conjunction with TPDL. The combination of the two conferences makes sense through the intersection of communities and subjects which bring mutual benefits. I hope that everyone present will take advantage of the opportunities that result from this co-organization that the University of Porto generously provides us.

DC-2018 features an exciting program consisting of articles and presentations, special sessions, workshops and a very special meeting: the DCMI Open Community Meeting. Why is this meeting so special? Because its nature faithfully translates the spirit of DCMI as an open, innovative and dynamic organization. Therefore, I invite those who view themselves as belonging to the DCMI community to participate in this meeting, but also those who do not yet know whether they belong or want to belong. All are welcome.

In the current year the female gender is in evidence in our conference and at DCMI. This evidence is shown by the fact that, for the first time, the chair of DCMI is a woman, but also by the fact that, for the first time, we have two women as keynote speakers. This is a signal given by all of us, women and men related to DCMI, that gender equity is important to DCMI and its events.

DCMI owes its name to the place of the first meeting that gave rise to its most iconic set of terms, and, on the way, to DCMI itself: the headquarters of the Online Computer Library Center in Dublin, Ohio, USA. From that date until today various developments and innovations related to metadata have been made by people linked to DCMI. Currently, everybody talks about Linked Data, but for us, at DCMI, it has been a long time since the data only makes sense if linked. And data is everywhere, not just in libraries and archives. Let us look around: where there is no data? This is the ultimate challenge that I leave you: to open up our horizons even further, to look for new paths and new ways of applying and sharing the solid knowledge and good practices that we have developed and will continue to develop over time.

Ana Alice Baptista, Chair, DCMI Governing Board

Program Committee Chairs' Welcome

Open Metadata for Open Knowledge is the overall topic of DC-2018. According to the Open Definition, knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness. Knowledge resources, be they data, text, or any other form of media, need metadata: to access the knowledge, to understand how to use it, to preserve its provenance and openness. Metadata is the basis for open knowledge and like the knowledge resources, it needs to be open and interoperable; to allow the exchange of knowledge and to create new ways to access and combine knowledge.

This year, the conference is co-located with the TPDL-2018, the 22nd International Conference on Theory and Practice of Digital Libraries. We believe this is an ideal constellation to discuss and collaborate on not only the idea of Open Knowledge, but also the necessary infrastructural developments that are required to provide open, trusted knowledge to everyone.

We are very pleased to present a program that reflects many current topics in the field of metadata and vocabularies. Metadata practitioners provide insights into their recent projects, with old and new challenges and chances, ranging from multilinguality to metadata quality. Wikidata gets more and more attractive not only as a data source, but also as central - critical? - infrastructure for open data. In good Dublin Core tradition, a variety of domains and applications is covered, ranging from smart cities over cultural heritage to social sciences and education.

Dublin Core conferences always have been first and foremost meetings of the metadata community, with lots of opportunities to engage in discussions and work together on current challenges. Consequently, we have this year again many special sessions in parallel to the presentations, to discuss topics like persistent identifiers, sustainable smart cities, or necessary changes in library and information science curricula.

We look forward to welcoming all participants and expect many fruitful discussions and encounters at DC-2018 in Porto.

Mariana Malta, CEOS.PP, Polytechnic of Oporto, Portugal

Kai Eckert, Stuttgart Media University, Germany

ORGANIZING COMMITTEE

Conference Co-Chairs

Ana Alice Baptista, Universidade do Minho, Portugal

Paul Walk, Dublin Core Metadata Initiative (DCMI) & Antleaf, Ltd., United Kingdom

Program Committee Co-Chairs

Kai Eckert, Stuttgart Media University, Germany

Mariana Curado Malta, CEOS.PP - Polytechnic of Oporto, Portugal

Program Committee

Akira Maeda, Ritsumeikan University, Japan

Ana Alice Baptista, Universidade do Minho, Portugal

Anne Gilliland, Department of Information Studies, UCLA, USA

Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Netherlands

Barbara Bushman, National Library of Medicine, USA

Bernhard Schandl, Gnowsis.com, Austria

Carol Jean Godby, OCLC, USA

Corey A. Harper, Elsevier Labs, USA

Cristina Pattuelli, Pratt Institute, USA

Corine Deliot, British Library, United Kingdom

Deborah Maron, UNC Chapel Hill, USA

Dion Goh, Nanyang Technological University, Singapore

Douglas Tudhope, University of Glamorgan, United Kingdom

Eero Hyvönen, Aalto University, Finland

Emmanuelle Bermes, Bibliothèque Nationale de France

Eva M. Méndez, University Carlos III of Madrid, Spain

Filiberto Felipe Martinez-Arellano, National Autonomus University of Mexico, Mexico

Hannah Tarver, University of North Texas Libraries, United States

Jacques Ducloy, University of Lorraine, France

Jess Peterson, Amazon, USA

Jian Qin, Syracuse University, USA

Jin-Cheon Na, Nanyang Technological University, Singapore

Johann Wanja Schaible, GESIS - Leibniz-Institute for the Social Sciences, Germany

Joseph A. Busch, Taxonomy Strategies, USA

Kai Eckert, Stuttgart Media University, Germany

Kevin Ford, Art Institute of Chicago, USA

Leif Andresen, Royal Danish Library, Denmark

Magnus Pfeffer, Stuttgart Media University, Germany

Makx Dekkers, Independent Consultant, Spain

Marcia Lei Zeng, Kent State University, United States

Mariana Curado Malta, CEOS.PP - Polytechnic of Oporto, Portugal

Mark A. Matienzo, Stanford University Libraries, USA

Masahide Kanzaki, Keio University Xenon Limited Partners, Japan

Michael K. Bergman, Structured Dynamics LLC**,** USA

Paul Walk, Dublin Core Metadata Initiative (DCMI) & Antleaf, Ltd., United Kingdom

Peter E Murray, Index Data, USA

Philipp Mayr, GESIS - Leibniz Institute for the Social Sciences, Germany

Ross Singer, Talis, USA

Ryan Shaw, University of North Carolina at Chapel Hill, USA

Shawne Miksa, University of North Texas, USA

Steven J. Miller, University of Wisconsin-Milwaukee, School of Information Studies, USA

Sarah Potvin, Texas A&M University Libraries, USA

Shigeo Sugimoto, University of Tsukuba, Japan

Stefanie Rühle, SUB Goettingen, Germany

Stuart A. Sutton, University of Washington, USA

Susanna Peruginelli, Susanna Peruginelli Library consultancy, Italy

Thomas Baker, Dublin Core Metadata Initiative (DCMI), Germany

Tomi Kauppinen, Aalto University, Finland

Vivien Petras, Humboldt-Universität zu Berlin, Germany

Wouter Klapwijk, Stellenbosch University, South Africa

Yi-Yun Cheng, School of Information Sciences, Univ. of Illinois at Urbana-Champaign, USA

TABLE OF CONTENTS

SESSION 5

61 **Validation**

SESSION 6

77 **Application Profiles**

SESSION 7

85 **Models**

107 POSTERS

AUTHOR INDEX

SESSION 1:
RDF

Linking knowledge organization systems via Wikidata
Joachim Neubert

An Approach to Enabling RDF Data in Querying to Invoke REST API for Complex Calculating
Xianming Zhang

Experiments in Operationalizing Metadata Quality Interfaces: A Case Study at the University of North Texas Libraries
Mark Edward Phillips & Hannah Tarver

Linking knowledge organization systems via Wikidata

Presentation

Joachim Neubert
ZBW – Leibniz Information
Centre for Economics,
Germany
j.neubert@zbw.eu

Keywords::Wikidata; STW; KOS; LOD; linked open data; alignment; matching tool

Abstract

Wikidata is a large collaboratively curated knowledge base, which connects all of the roughly 300 Wikipedia projects in different languages and provides common data for them. Its items also link to more than 1500 different sources of authority information. Wikidata can therefore serve as a linking hub for the authorities and knowledge organization systems represented by these "external identifiers". In the past, this approach has been applied successfully to rather straight-forward cases such as personal name authorities[1]. Knowledge organization systems with abstract concepts are more challenging due to, e.g., partial overlaps in meaning and different granularities of concepts.

Our work is based on the ongoing mapping effort of "STW Thesaurus for Economics" to Wikidata[2]. Contrary to other vocabularies, and just like Wikipedia, Wikidata can be extended by everybody and for every domain of human know knowledge. I will discuss the pros and cons of such extensions with regard to Wikidata itself and also to the indirect linking to Wikipedia pages, which is often one of the goals of mapping approaches to Wikidata. As an alternative to creating new Wikidata items, the newly introduced "mapping relation type" qualifier[3] – which comprises the SKOS mapping relations – allows for in-exact mappings of Wikidata items to external identifiers.

During mapping creation, in particular Wikidata's "Mix'n'match" tool and tailored SPARQL queries have proved useful. Existing mappings of other KOS to Wikidata can be exploited for deriving indirect mappings to these vocabularies, but also for generating mapping suggestions where a direct mapping already exists (as evaluated for a sample set of STW/GND mappings).

The extensibility of Wikidata by everybody also raises maintenance issues, as there is no single ownership and responsibility for a mapping. Wikidata's SPARQL query service allows tracing inconsistencies with according reports, taking into account the mapping relation types[4]. Coverage of newly introduced Wikidata and external concepts over time and possibly differing practices and interests of contributing parties pose additional challenges for the long-term maintenance of collaboratively used and curated vocabulary mappings. Besides technical challenges, this requires an understanding of Wikidata's policies and the communication within its community.

[1] Joachim Neubert: Wikidata as authority linking hub: Connecting RePEc and GND researcher identifiers, ZBW Labs 2017-11-30, http://zbw.eu/labs/en/blog/wikidata-as-authority-linking-hub-connecting-repec-and-gnd-researcher-identifiers

[2] https://github.com/zbw/stw-mappings

[3] https://www.wikidata.org/wiki/Property:P4390

[4]https://www.wikidata.org/wiki/Property_talk:P3911#Reports_for_the_maintenance_of_the_STW_ID_.2F_Wikidata_mapping

An approach to enabling RDF data in querying to invoke REST API for complex calculating

Xianming Zhang
Aviation Industry Development Center of China
forzxm@163.com

Abstract

RDF does not have very good support for calculation, especially complex calculation. SPARQL Inferencing Notation (SPIN) has been proposed with a specific capability of returning a value by executing external JavaScript file that in partly performs complex calculating, however it is still far away from accomplishing many practices. This paper investigates SPIN's capability of executing JavaScript, namely SPINx framework, presents a method of equipping RDF data with a new capability of invoking REST API, by which a user who is querying can obtain returned value by invoking the REST API performing complex calculating ,and then the value is semantically annotated for further use. Calculation of lift coefficient of airfoil is taken as a use case, in which with a given attack angle as input a desired returned value is obtained by invoking a particular REST API while querying the RDF data. Through this use case, it is explicit that RDF data invoking REST API for complex calculating is feasible and profound in both real practice and semantic web.

Keywords: spin, sparql, rdf, rest api

1. Introduction

In past years, a large number of RDF data and RDF-based applications have been developed for various domains. In order to take advantage of semantic feature of RDF, several query and rule languages, such as SPARQL[1], Jena rule (Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A. & Wilkinson, K., 2016) and SWRL (Horrocks, Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B. & Dean, M., 2004), have been developed and adopted widely. With these languages users can freely both query and reason about desired information from RDF data, and these languages also provide calculating capability by means of a number of inbuilt functions[2] for users to perform various calculations during either querying or reasoning.

Unfortunately, due to calculation complexity in real world, the calculating capability above is insufficient to accomplish many calculating tasks, the causes are as follows.

- Currently these inbuilt functions fail to perform many complex calculations such as matrix calculation, linear operation, these calculations are essential in such domain as physics and mechanics.

[1] See http://www.w3.org/TR/rdf-sparql-query and https://www.w3.org/Submission/SPARQL-Update - accessed August 23, 2018.

[2] See https://github.com/dotnetrdf/dotnetrdf/wiki/DeveloperGuide-SPARQL-XPath-Functions and http://jena.apache.org/documentation/query/library-function.html - accessed August 23, 2018.

- It is not wise to write too many calculating steps in both query and rule statements, as the feature of the statement is concise and explicit and too many calculating steps often harm this feature.
- Many calculations require external data as source, such as today's temperature for travel decision of today or exchange rate for economic decision, and the existing data either has too large volume to be appended into the new RDF dataset (every high cost and time consumption) or are commercial confidential not to be shared freely by others. Inbuilt functions provided by the languages above only consume data in the related RDF data.

In order to solve such problems this paper turns to SPARQL Inferencing Notation (SPIN)[3]. SPIN is the de-facto industry standard to represent SPARQL in form of RDF, and has been developed out of the necessity to perform calculations on property values. To be pertinent to this paper, SPIN provides a special framework (SPINx) that allows user-defined function to link an external JavaScript file to RDF data by RDF property, performs this user-defined function for calculation by invoking this linked JavaScript file and the resultant value can be semantically annotated by vocabulary of this RDF data. Of course in this situation, the content of linked JavaScript file is simple without many additional functionalities appearing in such working environment as web browser, so that the induced calculating capability is insufficient. But this paper investigates the mechanism of SPINx framework and devises a method to link REST API with RDF data, and invoke REST API while either querying or reasoning. It must be pointed out that this paper opens a profound ground in semantic web technology.

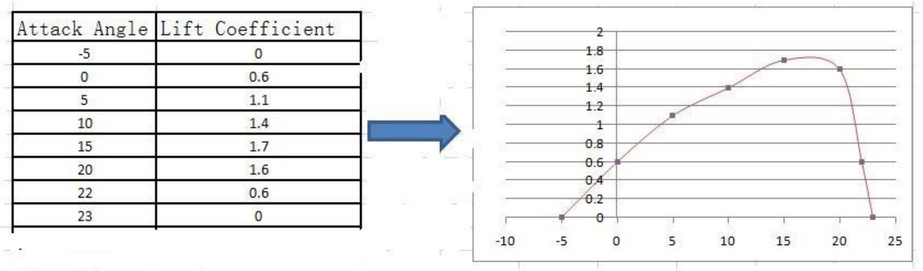
The rest of this paper is as follows: Section 2 presents a use case as motivation to illustrate it is both difficult and useful for a RDF data to deal with complex calculating; Section 3 introduces SPINx framework, especially the working principle of user-defined function linked with an external JavaScript file, and REST API[4]; Section 4 presents a solution to the use case by means the use of a SPINx framework to link with REST API; Section 5 presents the conclusions.

2 Motivation, a Use Case of Calculation for Lift Coefficient of Airfoil

A research group (here called Group A) builds a knowledge base on airfoil, which stores RDF data on airfoil, mainly explicit ones such as airfoil area and shape. An airfoil is designed to provide lift for airplane during flight, so it is necessary for users to query lift coefficient provided by a particular airfoil under a given circumstance. The lift-coefficient formula is as follow:

lift-coefficient=f(attack-angle) (F1)

In F1 there is no explicit formula (or calculation script) to accurately calculate lift coefficient from attack angle. Typically lift coefficient is a list of data through a limited number of experiments that record the data under different attack angles, as shown in Figure 1.



Attack Angle	Lift Coefficient
-5	0
0	0.6
5	1.1
10	1.4
15	1.7
20	1.6
22	0.6
23	0

FIG. 1. The table shows a list of attack angles and corresponding lift coefficient after a series of experiments, the graphic presents the curve graph build using the information of the table

[3] See https://www.topquadrant.com/technology/sparql-rules-spin. – accessed August 23, 2018

[4] See https://en.wikipedia.org/wiki/Web_API - accessed August 23, 2018

These experiments are conducted by another research group (here called Group B). The experimental data is locally stored in a web server controlled by Group B and is not freely accessible to others. As a result, the use and maintenance of this knowledge base faces the following problems:

- At present, a user can obtain either a list of data or a curve graph from this knowledge base, and fails to accurately know the lift coefficient of attack angles not shown in the data list, for example attack angles are 2,6,8 etc.
- The experiment is repeated for many times, new experimental data is appended and much of the existing data is modified. Though there is a lasting task of upgrading the knowledge base. The communication cost between Group A and Group B has to be taken into account.
- Group B does not deliver all of the data to this knowledge base for keeping others from getting the comprehensive sense of the experiment. In contrast, Group B allows others to query only one value of lift coefficient each time.

There is a solution to this problem, Group B can develop and deploy a REST API on the web server accessible to others. This REST API implements numerical approximation, such as least square and interpolating to cater for the first one problem, and retrieves data locally stored in web server to cater for the rest of problems. By invoking the REST API, users can obtain the desired results with attack angles and make use of the result for further reasoning. Now devising a unique method to link this knowledge base with REST API and to invoke it in querying (using SPARQL) is the prominent task.

3. Introduction to SPINx, the framework of SPIN for executing JavaScript

3.1 Brief Instruction to SPIN and SPINx

SPIN (SPARQL Inferencing Notation) queries are stored in RDF format and together with RDF data, which makes it possible to share SPARQL queries and update operations with other RDF data. A basic idea of SPIN is to link ontology classes in RDF with SPARQL queries that define constraints and rules formalizing the expected behavior of class members (instances). SPIN has become the de-facto industry standard to represent SPARQL rules and constraints on Semantic Web models and provides meta-modeling capabilities that allow users to define their own SPARQL functions, namely user-defined functions[5].

These user-defined SPIN function are very powerful ways of extending SPARQL, but they are still limited by whatever features are natively supported by the executing SPARQL engine. The SPINx framework included in SPIN makes it possible to define new SPARQL functions that are backed by JavaScript code. Whenever such new functions are invoked, a SPINx-aware SPARQL engine can look up the function's body and execute it using a JavaScript interpreter[6].

3.2 Calculation of the Use Case by means of SPINx Framework

Figure 2 illustrates the working principle of SPINx framework using the concrete use-case of the calculation of lift coefficient. The steps of FIG 2 are described in order as follows.

- Submitting SPARQL and SPINx framework loading RDF data

[5] See https://www.topquadrant.com/technology/sparql-rules-spin - accessed August 23, 2018

[6] See http://spinrdf.org/spinx.html – accessed August 23, 2018

A user submits a SPARQL statement into SPINx framework, the statement's goal is to calculate lift coefficient with Airfoil: CacuLiftCoefficient being the user-defined function (an instance of spin: Function) and 9 being the input attack angle. After reception the framework starts to search Airfoil: CacuLiftCoefficient in RDF data, and finds a desired segment that is compatible with the statement. Table 1 presents this situation.

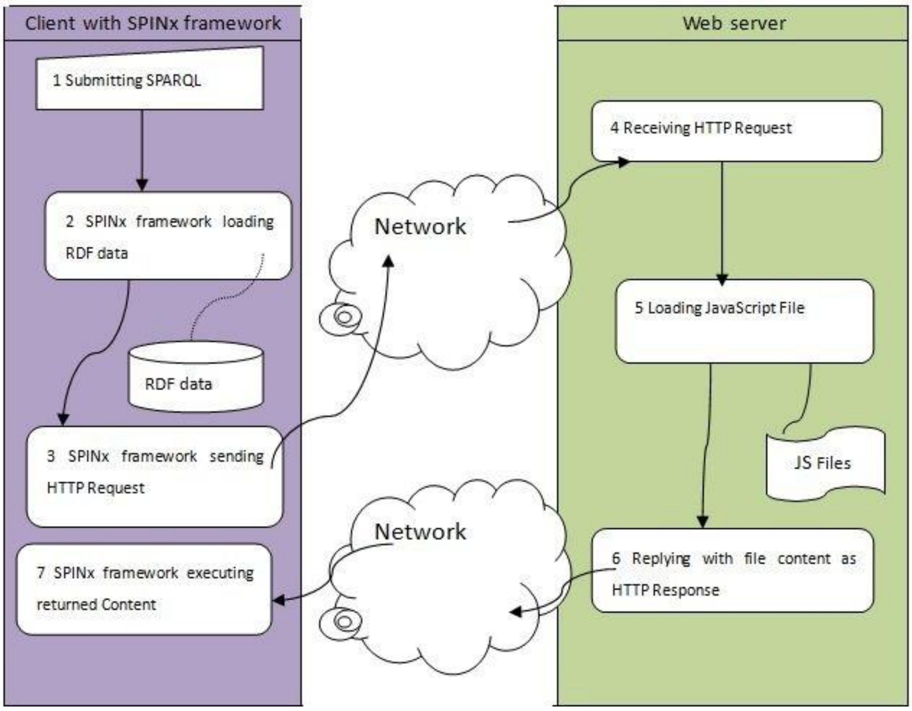

Client with SPINx framework
Web server
1 Submitting SPARQL
2 SPINx framework loading RDF data
RDF data
3 SPINx framework sending HTTP Request
7 SPINx framework executing returned Content
Network
Network
4 Receiving HTTP Request
5 Loading JavaScript File
JS Files
6 Replying with file content as HTTP Response

Fig. 1. Working principle of SPINx framework for the use case

TABLE 1: Mapping between RDF data and SPARQL statement

Segment of RDF Data	SPARQL Statement

Airfoil:CacuLiftCoefficient
a spin:Function ;
rdfs:subClassOf spin:Function ;
spin:constraint
[rdf:type spl:Argument ;
rdfs:comment "angle attack";
spl:predicate sp:arg1 ;
spl:valueType rdfs:Literal
] ;
spin:returnType xsd:float;
spinx:javaScriptFile "http://web-server/js/calculation.js"
	SELECT ?VALUE
	WHERE {
	BIND(Airfoil:CacuLiftCoefficient(9) AS ?VALUE).
	}

The spin: function Airfoil: CacuLiftCoefficient in the left column is mapped into Airfoil: CacuLiftCoefficient in the right, spl: Argument is mapped into 9, spin: returnType being mapped into ?VALUE means the returned data type is xsd: float, and the URL http://web-server/js/calculation.js is the implementation file of this function.

- SPINx framework sending and web server receiving HTTP request Submitting

After extraction of the URL of JavaScript file, namely http://web-server/js/calculation.js, the SPINx framework automatically sends a HTTP request to a web server identified in the URL. If the web server is connected to the sender/client via network by means of HTTP, it will smoothly receive this request. The excerpt of HTTP request is shown below.

```
[Request-Line]  GET /js/calculation.js

Host          web-server

Accept        text/html,*

Connection     keep-alive
```

- Loading JavaScript file and replying with its Content as HTTP Response

After analyzing the request correctly, the web server automatically retrieves its own file system for the JavaScript file calculation.js, reads content of this file and replies with it as HTTP response to the client.

```
HTTP/1.1 200 OK

Content-Type: application/x-javascript


function CacuLiftCoefficient(arg)

{

// for the sake of brevity, the code is omitted.

return result;

}
```

- SPINx framework executing returned content

After reception of HTTP response, SPINx framework extracts out the function body, semantically reorganize it with the segment of RDF data as well as the submitted SPARQL statement, and finally produces a piece of JavaScript code shown as following that in turn returns resultant value to user after being executed by SPINx framework.

```
return CacuLiftCoefficient(9);
```

3.3 Defect of the use of SPINx framework in the use case

It is impractical to include dataset in the files, especially the volume is not small and not allowed to expose freely.

In order to develop and deploy such JavaScript file, experimental data of confidential will have to be included in this exposed file and be updated frequently for future experiments, which both violates law of information security and is too expensive to maintain the JavaScript file.
As a result, most of applications of SPIN are focused on check constraints, perform data validation and some simple calculation accomplished by user-defined functions within RDF data (Furber & Hepp (2015); Riain & Mccrae (2012); Callahan & Michel (2012); Homola & Serafini (2012)), and its SPINx framework seemingly fails to play its deserved role in SPIN-related applications.

4. Driving SPINx framework to invoke REST API

Although it is impractical for researchers to develop a JavaScript file accessible on the web, but we can turn to an innovative method that is to develop a REST API that returns a small piece of JavaScript code containing resultant value after running. SPINx framework receives and executes this section of code quickly, namely extraction of resultant value in the code.

4.1 Introduction to REST API for the Use Case

A Web API (Application Programming Interface) is typically a defined set of HTTP request messages along with a definition of the structure of response messages for system-to-system interactions (information exchange programmatically)[7]. In this use case, the implementation of REST API is RESTful since it is popular with more and more web applications that have deployed their own REST APIs. Users can access and invoke REST API by means of http://web-server/REST/CacuLiftCoefficient/{AttackAngle}, where {AttackAngle} can be replaced with any real number, such as 9, 9.6 and so on. The REST API for calculating of lift coefficient, developed in Spring Boot, can be written as below

[7] See https://en.wikipedia.org/wiki/Web_API - accessed August 23, 2018

```
@RestController

@RequestMapping(path="/REST")

public class Calculator {

@RequestMapping(Path="/CacuLiftCoefficient/{AttackAngle}"

,produce=MediaType.TEXT_HTML_VALUE)

@ResponseBody

    public  String   CacuLiftCoefficient(@PathVariable("AttackAngle")
float arg1) {

           String result;

     float value



    // for sake of brevity, details are  omitted

    result="function           CacuLiftCoefficient()           {return
"+(String)value+";}";

    return result;

    }



}
```

After accessing to the REST API by URL, client can obtain a piece of code as HTTP response. For example, with http://web-server/REST/CacuLiftCoefficient/9 what client can obtain as follows:

```
function CacuLiftCoefficient () {return 1.34 ;}
```

4.2 Reconstructing working principle of SPINx framework with REST API

After reconstruction, the working principle is as below and for the sake of brevity, just steps in white box are addressed here.

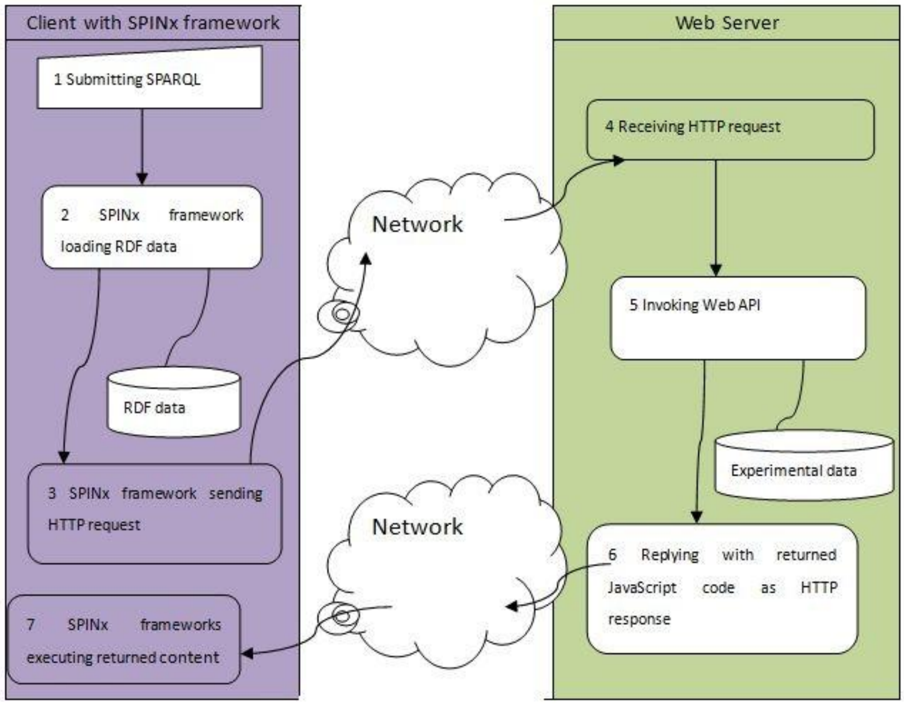

Client with SPINx framework
Web Server
1 Submitting SPARQL
2 SPINx framework loading RDF data
RDF data
3 SPINx framework sending HTTP request
7 SPINx frameworks executing returned content
Network
Network
4 Receiving HTTP request
5 Invoking Web API
Experimental data
6 Replying with returned JavaScript code as HTTP response

Fig. 1. Reconstructing the working principle of SPINx framework for the use case.

- Submitting SPARQL and SPINx framework loading RDF data

A new spin:Function instance called as Airfoil:CreateURL is added into RDF data, with which URLs for REST API of the use case can dynamically be created with variables of attack angle as input. This user-defined function is shown as below

```
Airfoil:CreateURL

      a spin:Function ;

      rdfs:label "create REST API URL"^^xsd:string ;

      rdfs:subClassOf spin:Functions ;

    spin:constraint

              [ rdf:type spl:Argument ;

                  rdfs:comment "angle attack";

                spl:predicate sp:arg1 ;

                spl:valueType rdfs: Literal

              ] ;

spin:returnType xsd:string;

spinx:javaScriptCode          "return          '          http://web-
server/REST/CacuLiftCoefficient/'+arg1"
```

A user submits SPARQL statement as below before actually beginning to query resultant value for a given input, in which the number 9 is the given input. It is noted that spinx: javaScriptCode links a piece of JavaScript code that be executed locally in querying.

```
DELETE
 { Airfoil:CacuLiftCoefficient  spinx:javaScriptFile ?OLDURL
 }
 INSERT
 { Airfoil:CacuLiftCoefficient  spinx:javaScriptFile ?NEWURL
 }
WHERE{
Airfoil:CacuLiftCoefficient  spinx:javaScriptFile ?OLDURL.
BIND(Airfoil:CreateURL(9) AS ?NEWURL ) }
```

Now the user can query resultant value by submitting SELECT ?value WHERE {BIND(Airfoil:CacuLiftCoefficient( ) AS ?value). }

- Invoking REST API and replying

After analyzing the request correctly, the web server automatically finds and invokes Calculator. CacuLiftCoefficient (9) ,the REST API that organizes the 9 with the experimental data in either data file or database, and processes them with a specific algorithm. After processing, returned value is replied as HTTP response to the client.

```
HTTP/1.1 200 OK
Content-Type: application/x-javascript
function CacuLiftCoefficient(){return 1.34;}
```

It should be noted that such SPARQL operations occur separately, which means the resultant value are independent from each other. Through this example, the feasibility of invoking REST API with SPARQL statements is approved.

5. Conclusions

With abundance in IT infrastructure today, the number of REST API is growing and RDF-based knowledge system should be constructed by fully taking advantage of this situation including REST API rather than from scratch. This paper discusses that usefulness and feasibility of using SPINx framework to invoke REST API for resultant value. It can be said that the paper's achievement is a breakthrough in development of RDF-based knowledge system. In my opinion, the study of this paper can make the semantic web models obtain powerful calculating capacity.

Of course in such situations, the coordinating asynchronous requests, latency, availability and security must be taken into account, these problems should be solved effectively (at least in part) as the technologies for REST API, exemplified by SPRING BOOT, has made much effort to solve them from birth. Many readers familiar with REST API may put forward such viewpoint that majority of REST APIs available are not intended to return a piece of JavaScript code. To solve this problem is to establish a proxy as intermediate between clients and web servers with HTTP as communication protocol. Clients send HTTP request to specific REST APIs on the proxy and the specific REST APIs request normal REST APIs on web servers. In return, the resultant value will be wrapped in a piece of JavaScript code by the proxy and then send back to clients.

Acknowledgements

I would like to thank the reviewers of this article who, with their interest and commitment, contributed to its clarification and improvement.

I would like to thank Mariana Curado Malta and Paul Walk. Since English is not my native language and it is the first time I submit an article to an international conference, I encountered difficulties. They arranged a zoom meeting for me to talk about the conference, Mariana instructed and helped me to typeset this paper, and to reference using the APA style. Paul cleaned my paper for a better understanding. Especially Mariana paid much painstaking care for my paper, and I can not imagine the result without her care. I am truly grateful for their help.

References

Callahan, Alison & Michel Dumontier (2012). Evaluating scientific hypotheses using the SPARQL Inferencing Notation. The Semantic Web: Research and Applications, Lecture Notes in Computer Science 7295, 647-658.

Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A. & Wilkinson, K. (2016). The Jena Semantic Web Platform: Architecture and Design. In HP Laboratories Technical Report HPL-2003-146.

Furber, Christian & Hepp, Martin. (2015). Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In Business Information Systems, Lecture Notes in Business Information Processing, vol. 47, 35-46.

Homola, Martin & Serafini, Luciano. (2012). Contextualized knowledge repositories for the semantic. Journal of Web Semantics , Vol 12, 64-87

Horrocks, Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B. & Dean, M. (2004). Swrl: A semantic web rule language combining owl and rule ml. W3C Member Submission, Retrieved August 23, 2018, from http://www.w3.org/Submission/SWRL.

Riain, Sean O & Mccrae, John P. (2012). Using SPIN to Formalize Accounting Regulations on the Semantic Web. In ESWC 2012: The Semantic Web: ESWC 2012 Satellite Events, 58-72.

Experiments in Operationalizing Metadata Quality Interfaces: A Case Study at the University of North Texas Libraries

Mark Edward Phillips
University of North Texas Libraries,
United States
mark.phillips@unt.edu

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

Abstract

This case study presents work underway at the University of North Texas (UNT) Libraries to design and implement interfaces and tools for analyzing metadata quality in their local metadata editing environment. It discusses the rationale for including these kinds of tools in locally-developed systems and discusses several interfaces currently being used at UNT to improve the quality of metadata managed within the Digital Collections.

Keywords: metadata quality; user interfaces; metadata quality interfaces; Web interfaces

1. Introduction

Digital collections in cultural heritage institutions including libraries, archives, museums and galleries have grown steadily over the past decade. As technologies for the digitization of analog collections and the accumulation of born-digital materials has become more accessible to institutions of all sizes, these same institutions have made great efforts toward making digital resources available via the web. With this increase, many have begun to focus on the quality of the metadata that describe these resources. Analysis of metadata for digital resources has been conducted on large aggregations in the Digital Public Library of America (DPLA) in the United States, and Europeana in the European Union (Harper, 2016; Tarver, Phillips, Zavalina & Kizhakkethil, 2015). This work has led to discussion on how to communicate needed metadata improvements to local repositories (Dangerfield, 2015). While this remains an unsolved problem, there is another gap that is not as often discussed: mainly, how are local repositories experimenting with tools and interfaces to understand the quality of the metadata in their own systems, and how are these same tools and interfaces used in practice?

Of course, one concern for anyone working with metadata is determining the quality of the data, such as the existence of typos, missing or mislabeled information, or improper formatting. These errors can be introduced in a number of ways, including data input errors, importation of data that has different formatting, and values based on outdated rules. The larger the collection, the more difficult it can be to check for errors manually. The Digital Library Federation Assessment Interest Group Metadata Working Group has started to collect documentation and tools as a first step toward providing guidance for local repositories (DLF AIG MWG Metadata Assessment Toolkit, n.d.), though there is a wide range of needs.

It is almost expected today that there are tools and interfaces built into digital library platforms to help metadata editors assess and understand the quality of the metadata that they are creating. From our research we have not discovered that this is the case. One of the challenges that we see in this area of research is that most of the tools and interfaces that have been developed by institutions may be focused solely on their local situations, workflows, and data models, and therefore have not been broadly shared with others. This is unfortunate because there is much that we can learn from others related to what they are trying to accomplish, how they are working toward these goals, and the interfaces and systems that they are putting in place. This case study does not attempt to define or characterize specific quality measures in the Digital Collections, but it discusses the work underway at the UNT Libraries, focused on building tools and interfaces for reviewing and generally improving metadata.

1.1 Background

The University of North Texas (UNT) Libraries' Digital Collections comprise more than 2.2 million items, housed in a single administrative system and publicly accessible via three interfaces. The Portal to Texas History (https://texashistory.unt.edu/) contains materials owned by nearly 400 different partner institutions across the state of Texas; the UNT Digital Library (https://digital.library.unt.edu/) contains items owned, created, or licensed by UNT, including current scholarly works; the Gateway to Oklahoma History (https://gateway.okhistory.org/) contains items owned by the Oklahoma Historical Society. The level of collaboration across the Digital Collections means that a number of metadata editors work within the system to create or change metadata. Since 2009, more than 700 unique editors have edited records in the metadata editing system, including trained staff members, catalogers, library science students, and volunteers.

The current digital library system was developed in-house using open-source components. It was completed in 2009 and has undergone a number of iterative changes to both the public and administrative interfaces. Metadata in the Digital Collections is based on Dublin Core with the addition of local fields and qualifiers for a possible twenty-one fields used for all items in the system, including eight that are required for every record. There are extensive guidelines in place outlining the technical and semantic expectations for metadata in each field.

This paper seeks to discuss some of the experiments in tools and interfaces being developed at the University of North Texas Libraries that help metadata creators identify and improve deficiencies in their collections of metadata.

2. Analysis Tools

As the UNT Libraries' Digital Collections have grown, we have become increasingly aware that we need tools to allow us to understand the quality of the metadata that is being created in these collections and to analyze or compare larger and larger sets of data. The first tool, called the "Metadata Analysis Tool" was built in 2005; sadly, because its features were only used internally, there were only a few external presentations and no published discussion of how we used the tool in our systems. The Metadata Analysis Tool was forgotten in our library as we migrated our digital collections from system to system.

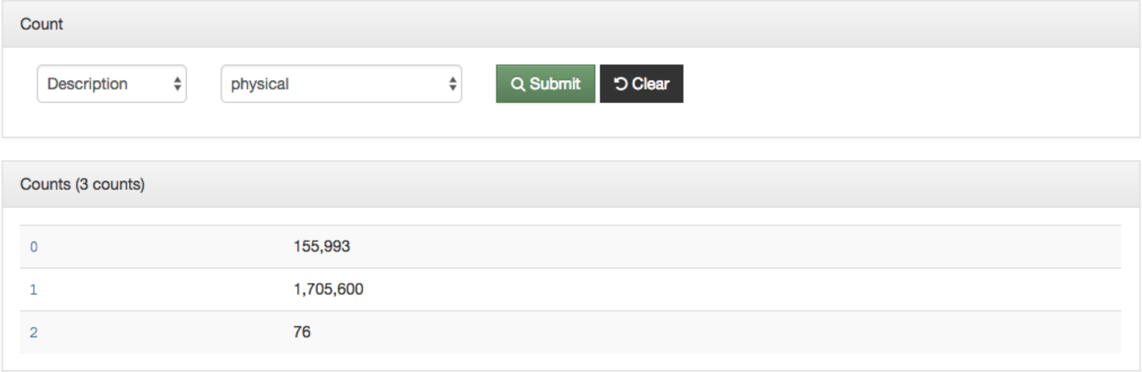
For a number of years, we have been doing some basic analysis on record values by harvesting the records and using Python scripts to look at field values (Phillips, 2013). Although this is useful, there are some downsides: it is not always easy to check values across multiple collections or the whole system; it can be difficult to check everything systematically without a particular concern in mind; and importantly, this method is not particularly accessible to the many editors working on metadata in our system. We wanted to move toward tools that could be used by metadata editors to check their own work, or to identify problems throughout the system and start correcting them.

Some institutions have had success with tools like OpenRefine for cleaning up metadata for their digital collections. We, too, have used OpenRefine for projects to improve metadata before it is added to our primary digital library platform (Phillips, Tarver, & Frakes, 2014). Like many, we found that OpenRefine is a wonderful tool for working with spreadsheets and other types of data, but there are a few challenges. First of all our data generally isn't rectangular and doesn't easily fit into a spreadsheet representation. We have some records with one creator and others with dozens of creators. There are ways to work with these kinds of data but it can get complicated. A bigger challenge we have in our local environment is that while many systems can generate a spreadsheet of their data for exporting, very few -- including our metadata management system -- have a way of importing those changes back into the system in a spreadsheet format. This means that while you can pull data from the system and clean it up in OpenRefine, there is no way to get that nice clean data back into the system. A way that we found that we could use OpenRefine was to identify records to change and then go back into the system and edit records there; however it is a tedious and time-consuming process. In order to overcome this set of challenges we decided that we needed to build analysis tools directly into the metadata-editing interface used for the Digital Collections. That way our metadata editors could identify a problem and immediately fix it in an interface they understand and use every day.

2.1 Facet and Count Interfaces

During summer 2017, our software development team implemented the first of our suite of integrated analysis tools: Count and Facet. For each of the tools -- including Cluster, which was added later and is described further in the next section -- an editor must choose a specific field but has the option, when applicable, to limit to any qualifier, to a specific qualifier, or to values that have no qualifier. Editors also have the ability to filter the record analysis based on other criteria, such as collection or institution, material type, public visibility, or records that the editor has modified. These criteria and results of the analysis assist in identifying obvious problems, such as records without specific types of required values or existing values that do not have qualifiers.

Count sifts records based on the number of entries in a field so that editors could see, for example, that there are 65,772 records containing 0 subject entries, 23,026 records containing 1 subject entry, 12 records containing 87 subject entries, etc. Figure 1 shows counts for physical description entries, for which records should never have multiple entries and, ideally, ought to have a single entry, though it is not required. Currently, 76 records have two description entries labeled “physical description” and nearly 156,000 records have no physical description. Based on these counts, an editor should review the 76 records with multiple physical descriptions to fix qualifiers (if values are mislabeled) and to move or collapse information as needed, to eliminate multiple entries. As a longer-term project, we would also want editors to start adding physical descriptions to the 156,000 records without values and to review those records, a many of them likely have other errors or omissions.



Count

Description: physical

Counts (3 counts)

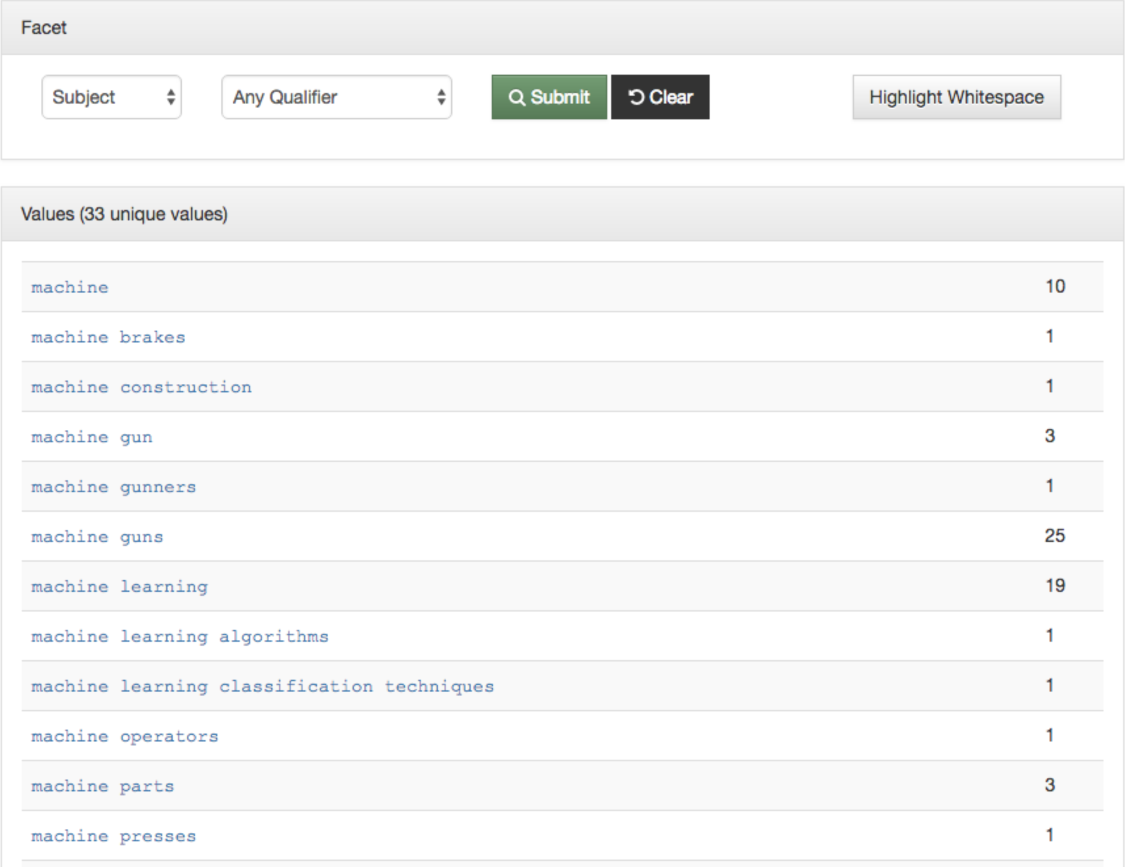
0	155,993
1	1,705,600
2	76

Figure 1. Example values in the “Count” tool for physical description entries.

Facet lists all of the unique values for a particular field and the frequency, i.e., the number of records in which each term appears (see Figure 2). This tool is most useful for finding typos and small inconsistencies across values in a field, such as “machine gun” versus “machine guns.” It also lets an editor see the most commonly used terms in a collection or across the system.

Values for qualifiers, some sub-fields, and five required fields (language, resource type, format, collection, and institution) are managed in local controlled vocabularies and are connected directly to the edit interface as drop-down menus to prevent non-valid terms. Though we also encourage the use of other sources of controlled vocabularies -- e.g., VIAF Virtual International Name File) or LCNAF (Library of Congress Name Authority Files) for names; LCSH (Library of Congress Subject Headings), Legislative Indexing Vocabulary (LIV), Chenhall's Nomenclature for Museum

Cataloging, etc. for subjects -- we don't currently have a way to validate against external controlled vocabularies and the terms are not always consistent. Facet can be useful in these instances to compare controlled and uncontrolled terms alphabetically to see where there might be overlap and to determine when it might be appropriate to change values.



Values (33 unique values)

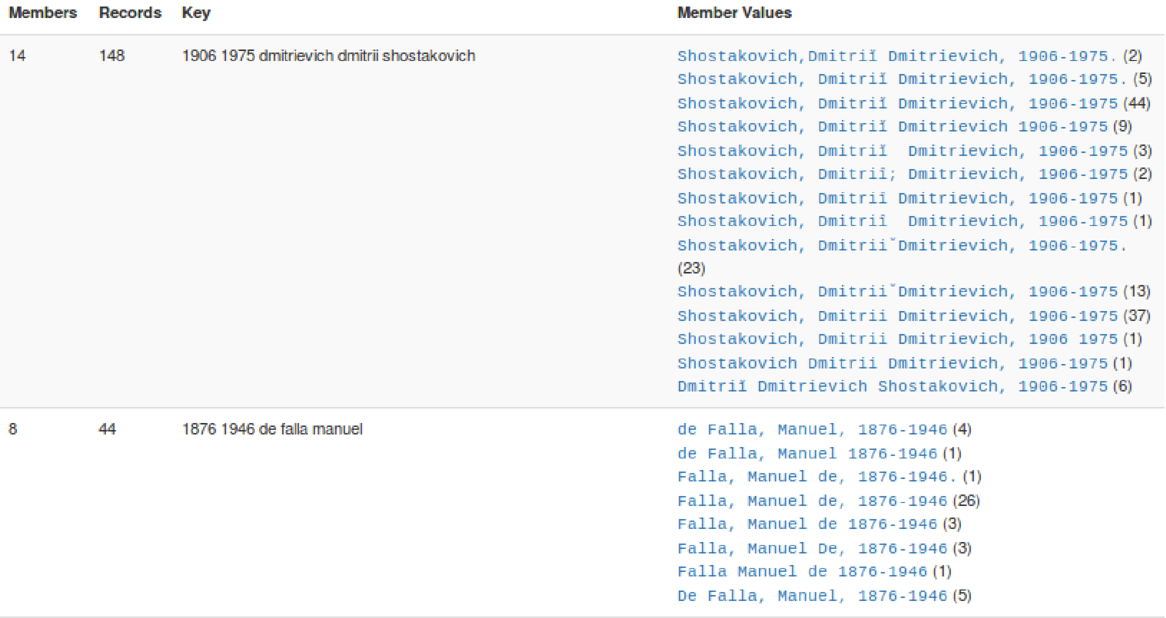
machine	10
machine brakes	1
machine construction	1
machine gun	3
machine gunners	1
machine guns	25
machine learning	19
machine learning algorithms	1
machine learning classification techniques	1
machine operators	1
machine parts	3
machine presses	1

Figure 2. Example values in the "Facet" tool for all subject entries.

2.2 Cluster Interface

The third tool that we introduced is more complicated and uses algorithms to normalize values and look for matches, the same way that OpenRefine can cluster possible matches in spreadsheet data. In this tool, a user has the same options, with the addition of a drop-down menu to choose an algorithm. Cluster is intended to automatically group together values that are most likely to be differently-formatted versions of the same term. Adjusting the algorithm used to normalize the values can result in different clusters, depending on the types of variations that an editor might want or expect to find (see Table 1).

Most of the time the default (fingerprint) algorithm is sufficient. Fingerprint normalizes the values by changing all characters to lowercase, simplifying non-ASCII characters, replacing punctuation with spaces, removing spaces at the start or end of the term, collapsing duplicate spaces within the term, alphabetizing the tokens, and deleting any duplicate tokens. Our implementation of this fingerprint algorithm is the same that is used by OpenRefine (Clustering in depth, n.d.). We have found it to be a good baseline algorithm for metadata editors (see Figure 3). Each cluster displays the number of members (unique values), the number of records containing the clustered values, the key (normalized text string), and the member values (existing values with the number of records in which they appear). Clusters can be sorted alphabetically based on the cluster key, by number of total records or members, by total length, and by the amount in variation of length among cluster members.



Members	Records	Key	Member Values
14	148	1906 1975 dmitrievich dmitrii shostakovich	Shostakovich,Dmitriĭ Dmitrievich, 1906-1975. (2) Shostakovich, Dmitriĭ Dmitrievich, 1906-1975. (5) Shostakovich, Dmitriĭ Dmitrievich, 1906-1975 (44) Shostakovich, Dmitriĭ Dmitrievich 1906-1975 (9) Shostakovich, Dmitriĭ Dmitrievich, 1906-1975 (3) Shostakovich, Dmitrii; Dmitrievich, 1906-1975 (2) Shostakovich, Dmitrii Dmitrievich, 1906-1975 (1) Shostakovich, Dmitrii Dmitrievich, 1906-1975 (1) Shostakovich, Dmitrii˘Dmitrievich, 1906-1975. (23) Shostakovich, Dmitrii˘Dmitrievich, 1906-1975 (13) Shostakovich, Dmitrii Dmitrievich, 1906-1975 (37) Shostakovich, Dmitrii Dmitrievich, 1906 1975 (1) Shostakovich Dmitrii Dmitrievich, 1906-1975 (1) Dmitriĭ Dmitrievich Shostakovich, 1906-1975 (6)
8	44	1876 1946 de falla manuel	de Falla, Manuel, 1876-1946 (4) de Falla, Manuel 1876-1946 (1) Falla, Manuel de, 1876-1946. (1) Falla, Manuel de, 1876-1946 (26) Falla, Manuel de 1876-1946 (3) Falla, Manuel De, 1876-1946 (3) Falla Manuel de 1876-1946 (1) De Falla, Manuel, 1876-1946 (5)

Figure 3. Example clusters for contributor names using the fingerprint algorithm.

Once the basic framework was in place for applying an algorithm to a string to perform normalization and hashing into buckets, we started to experiment with variations on algorithms that would be useful in specific circumstances, outlined in Table 1.

TABLE 1: Clustering algorithms with example values.

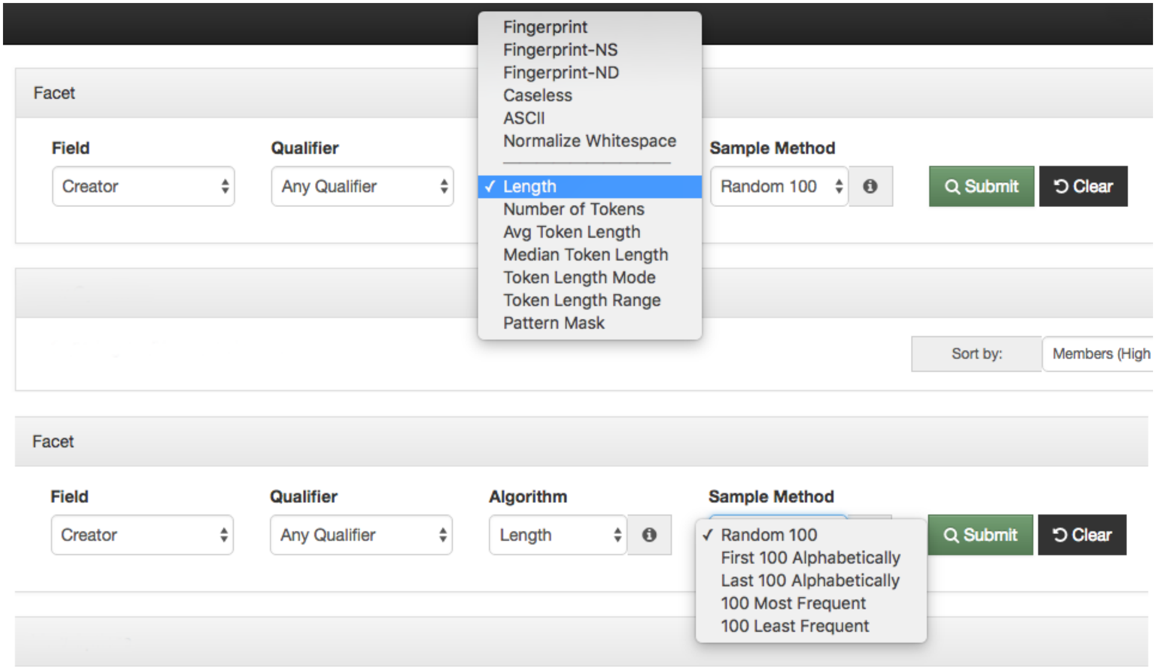
Algorithm	What it Does	Example Input	Example Output
Fingerprint	Normalizes capitalization & punctuation, deletes duplicate words	Wereszczak, Andrew A.	a andrew wereszczak
Fingerprint - No spaces	Same as fingerprint & removes punctuation without changing spacing	F.B.I.	fbi
Fingerprint - No dates	Same as fingerprint & ignores dates	Schmidt, Brian A., 1980-	a brian schmidt
Caseless	Makes all values lowercase	Austin, Stephen F.	austin, stephen f.
ASCII	Converts letters with diacritics to their plain ASCII representation	Castillo, José	Castillo, Jose
Normalize Whitespace	Replaces repeated whitespace with a single whitespace character	David S. Castle Co.	David S. Castle Co.

Alternative algorithms or customized versions of algorithms can also be added as needed, to isolate or eliminate particular kinds of values within clusters. We have noticed that some algorithms work better for certain fields; for example, the Fingerprint - No Dates algorithm works most effectively on the creator and contributor fields that contain many values that only differ by the inclusion of dates, such as authorized forms of names in the Library of Congress authority file versus unauthorized forms. This also works for numeric symbol codes. For example, the cluster for Shostakovich (Figure 3) gains a fifteenth member using the "no dates" version -- Shostakovich, Dmitrii&#774; Dmitrievich, 1906-1975 – and composer Gabriel Faure (not visible in the example) gains the variation "Faure, Gabriel" alongside "Faure&#769;, Gabriel, 1845-1924." and Faure, Gabriel, 1845-1924, among others.

2.3 Sampling

When working with the normalization algorithms described above, the goal was to identify groups or clusters that contain two or more values. If a cluster only had a single value, it was ignored and not displayed in the interface. We found that this was not always desirable, but when we wanted to analyze values in groups that contained a large number of members, we ran into problems with the interface and how to display these sets.

The Cluster tool has a useful framework to group values by specific features, such as length or alphanumeric patterns. For these cases, because every value is included in the results, some clusters get extremely large and would be prohibitive to display. Instead, clusters with over 100 values are displayed by sampling according to chosen criteria -- random values, first or last values alphabetically, most or least frequent -- so that each cluster is a reasonable size (see Figure 4).

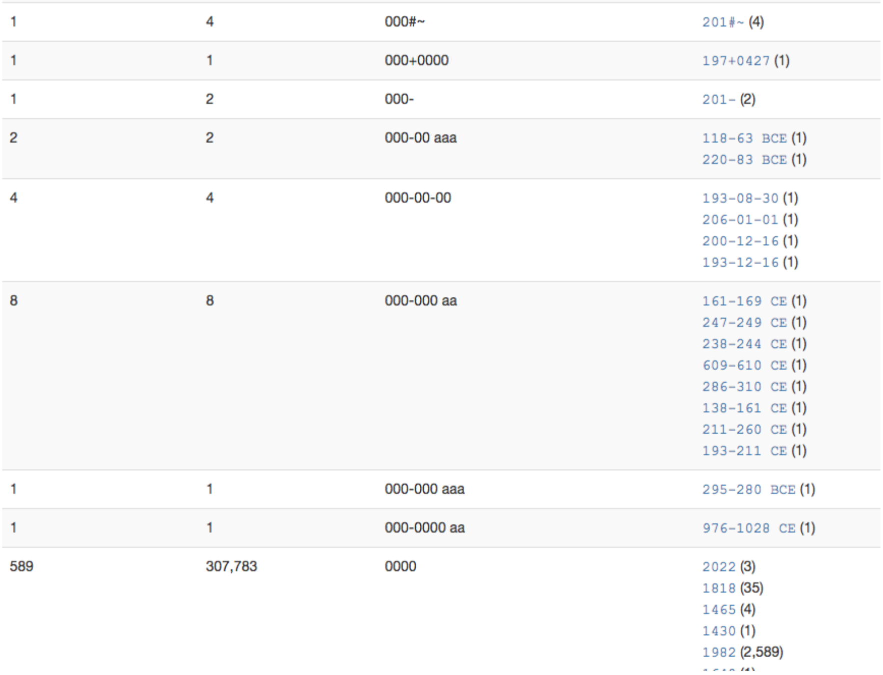


Field: Creator
Qualifier: Any Qualifier
Algorithm: Length
Sample Method: Random 100
Sort by: Members (High

Figure 4. Menu options in the "Cluster" tool for algorithms that use sampling.

These sampled algorithms assist in identifying values that are outliers, such as subject values that have only one character or that have more than 1000 characters. It also makes it easier to sort by patterns or lengths, which can be helpful for certain fields.

An algorithm that uses this sampling is our Pattern Mask algorithm that takes the input string and converts any digits to the character '0' and any letters to the character 'a', while leaving any punctuation as it is (see Figure 5). An example of this would convert the EDTF (Extended Date/Time Format) date of '194u' into the string '000a'. We aren't the first to find this especially useful for analyzing dates because it allows date values with similar patterns to group together, such as '193u' and '194u', which both convert to '000a' (Van Hooland, 2009).



1	4	000#~	201#~ (4)
1	1	000+0000	197+0427 (1)
1	2	000-	201- (2)
2	2	000-00 aaa	118-63 BCE (1) 220-83 BCE (1)
4	4	000-00-00	193-08-30 (1) 206-01-01 (1) 200-12-16 (1) 193-12-16 (1)
8	8	000-000 aa	161-169 CE (1) 247-249 CE (1) 238-244 CE (1) 609-610 CE (1) 286-310 CE (1) 138-161 CE (1) 211-260 CE (1) 193-211 CE (1)
1	1	000-000 aaa	295-280 BCE (1)
1	1	000-0000 aa	976-1028 CE (1)
589	307,783	0000	2022 (3) 1818 (35) 1465 (4) 1430 (1) 1982 (2,589)

Figure 5. Clusters for date entries using the pattern mask algorithm and random sampling.

The screenshot provided of the Pattern Mask algorithm demonstrates the need for the sampling mechanism, since '0000' is very common, containing 589 different values that represent 307,783 records. These 589 different values would take up considerable space on the page and generally are less likely to require investigation than clusters that contain only a few different values; instead only 100 random values are displayed.

3. Analysis Tools in Relation to Quality

Although the tools and interfaces cannot identify problems automatically, each of them is meant to assist editors in finding values that are incorrect or that may need to be verified for accuracy. In some cases, clustered values could all be correct. For example, strings that contain duplicate word tokens will cluster together using fingerprint -- e.g., *United States - Texas - Denton County* and *United States - Texas - Denton County - Denton* -- even though these may be separate, unique terms that are both valid.

However, these tools are still useful for improving the quality of records because they provide editors with relatively easy ways to analyze values in a particular collection to find typos, incorrect formatting, missing values, and other inconsistencies that would be time-consuming and difficult to identify through spot-checking or other methods of proofreading.

Since the Facet, Count, and Cluster interfaces are directly integrated into the editing system, results in the analysis tools connect to item records. Clicking on a count/value in any of the tools opens a new tab with the standard Dashboard displaying search results for only the records that have that count/value and that meet any other selected criteria. Although there is no way to make "batch" edits or "find-and-replace" across our system, this integration makes it relatively easy to identify specific, known problems and provide a link to an editor who can change the records or review them, as needed, to improve the consistency and overall quality.

3.1 Metrics

We have begun to use the number of clusters as a rough but relatively accurate metric for one aspect of quality; i.e., if the number of clusters in a collection is reduced, that would represent an increase in consistency and lower entropy, and therefore an increase in quality, at least for a particular field. For one specific project, a person was assigned to start fixing formatting problems with names in a technical report collection and to keep track of the number of clusters to show progress. What we actually found in that instance is that the number of clusters often increases before it decreases. For example, if records in the collection have J. D. Smith represented three different ways -- Smith, J.D.; Smith JD; and Smith, J D -- the first two versions would group together using the basic fingerprint algorithm, but the third would not be in the cluster. After some of those names are corrected to the version we prefer in our system (Smith, J. D.), the same algorithm would now cluster together the third original version and the new, correct version, doubling the number of clusters for that name. Since some of the individual records contain multiple authors, fixing all of the names in one record could add four or five new clusters in some cases, during the process of clearing one.

Essentially, while the number of clusters can still give a general sense of consistency (particularly for certain fields or within specific collections), using the number of clusters as a relatively exact metric, or expecting that it represents the actual number of corrections appears to be much fuzzier than we had first expected. Despite this we have found this metric useful over longer periods of time to show the improvement in consistency within a given field across a collection of metadata records. Similarly, the number that displays in each tool provides some reflection of work; for example, if the number of unique values decreases (as values are made more consistent) or increases (as missing values are added) when using Facet to analyze fields.

4. Discussion

As we mentioned above the benefit we have found with the approach of building these metadata quality interfaces into our digital library system is that it allows our metadata editors direct access to records that have an identified issue. Another benefit is that as metadata records change, the interfaces are quickly updated.

A limitation that we've found especially with the clustering interface is the amount of time it takes to generate the clusters across all two million records in our system. Because each clustering algorithm has to operate a string transformation on every unique value for a field, there is the possibility of over a million iterations for a large field like subject. With the current implementation some clusters take almost twenty seconds to generate. As we add more records with more values to the system this time will likely increase, causing users to wait for longer periods of time. To overcome some of the wait time, we currently cache clusters for ten minutes before they are regenerated. Although that means that the values do not update in real time as changes are made, this seems to be a reasonable tradeoff for users as they often work on a number of different clusters before needing to have the clusters regenerated.

The tools and interfaces discussed in this paper are useful in helping to identify problems, but they are a first step, both for future development in our system and for additional quality-related research. There may be ways to implement new tools or changes to existing interfaces depending on needs expressed by metadata editors or ways that certain fields affect the usability of public interfaces and usefulness to external users. One ongoing question is how to prioritize corrections. Looking at values across more than two million records results in an extremely large number of known problems and outliers that could be possible errors. There are many ways to organize that information. For example, is there more benefit in fixing clusters that affect a larger number of records, versus clusters that have a larger number of members (less consistency)?

This case study was designed to demonstrate some of the new tools and interfaces that have been developed for the UNT Libraries' Digital Collections. While there is a plan to eventually release the software for these interfaces, it is unlikely it would be adopted by other institutions because of the specific design decisions that were made to meet our local needs. What we do hope is that others who have thought about building metadata quality tools and interfaces will see this case study and will be interested in developing similar tools or interfaces in their local environments. Interfaces for metadata management are often only available to locally-authenticated users, so it is usually impossible to see locally-developed tools for working with metadata. By sharing this case study we hope that others are encouraged to share their work, both as code, but more importantly with discussion about how and why the tools and interfaces were developed in the first place.

5. References

Clustering in depth (n.d.). In GitHub. Retrieved May 7, 2018 from https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Dangerfield, Marie-Claire et. al (2015). Report and recommendations from the Task Force on Metadata Quality. Retrieved from http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf

DLF AIG MWG Metadata Assessment Toolkit (n.d.). In GitHub. Retrieved August 13, 2018 from http://dlfmetadataassessment.github.io/

Harper, Corey. A. (2016, June). Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). Code4Lib (33). Retrieved from http://journal.code4lib.org/articles/11752

Phillips, Mark Edward (2013, January). Metadata analysis at the command-line. Code4Lib (19). Retrieved from https://digital.library.unt.edu/ark:/67531/metadc157309/

Phillips, Mark Edward, Hannah Tarver & Stacey Frakes (2014, January). Implementing a collaborative workflow for metadata analysis, quality improvement, and mapping. Code4Lib (23). Retrieved from https://digital.library.unt.edu/ark:/67531/metadc284577/

Tarver, Hannah, Mark Phillips, Oksana Zavalina, & Priya Kizhakkethil (2015). An Exploratory Analysis of Subject Metadata in the Digital Public Library of America. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2015, 30-40

Van Hooland, Seth (2009). Metadata quality in the cultural heritage sector: Stakes, problems and solutions (Thesis). Retrieved from http://homepages.ulb.ac.be/~svhoolan/these.pdf

SESSION 2:
Multilingual

A study of multilingual semantic data integration
Douglas Tudhope & Ceri Binding

Designing a Multilingual Knowledge Graph as a Service for Cultural Heritage – Some Challenges and Solutions
Valentine Charles, Hugo Manguinhas, Antoine Isaac, Nuno Freire & Sergiu Gordea

A study of multilingual semantic data integration
Presentation

Douglas Tudhope
Hypermedia Research Group,
University of South Wales, UK
douglas.tudhope@southwales.ac.uk

Ceri Binding
Hypermedia Research Group,
University of South Wales, UK
ceri.binding@southwales.ac.uk

Keywords: knowledge organization systems; linked open data; query expansion; natural language processing; Getty Art & Architecture Thesaurus; CIDOC Conceptual Reference Model; archaeology;

Abstract

The availability of the various forms of open data today offers great opportunity for meta level research that draws on combinations of data previously considered only in isolation. There are also great challenges to be overcome; datasets may have different data models, may employ different terminology or languages, project data may only be represented by the final textual report. However, metadata and controlled vocabularies have the potential to help address many of these issues.

Previous work by the authors has explored semantic integration of English language archaeological datasets and reports (Binding et al., 2015; Tudhope et al., 2011). This presentation reflects on experience from a semantic integration exercise involving archaeological datasets and reports in different languages. Different forms of Knowledge Organization Systems (KOS) were key to the exercise. The Getty Art and Architecture Thesaurus (AAT) was used as the underlying value vocabulary and the CIDOC CRM ontology as the metadata element set (Isaac et al. 2011) for the semantic integration. Linked data expressions of the vocabularies formed part of an integration dataset (RDF) extracted from the source data, together with subject metadata automatically generated from the reports via Natural Language Processing (NLP) techniques.

The data was selected following a broad theme of wooden material, objects and samples dated via dendrochronological analysis. The investigation was conducted as an advanced data integration case study for the ARIADNE FP7 archaeological infrastructure project (ARIADNE 2017), with the datasets and reports provided by Dutch, English and Swedish ARIADNE project partners.

The presentation will outline the data cleansing, NLP and integration methods and present illustrative scenarios from the web application Demonstrator (2017). A template based tool was used for data conversion of extracts from the archaeological datasets and also the data resulting from NLP information extraction from the archaeological reports (STELETO 2016). Following the approach used in the ARIADNE Portal (2017), terms from different languages were intellectually mapped to concept identifiers from the Linked Open Data implementation of the Getty AAT (2018), in order to support cross search (via the AAT) over subject metadata in different languages. The user is shielded from some of the complexity of the metadata framework and the underlying SPARQL implementation by an interactive query builder. The search system exploits the AAT's hierarchical relationships and specialised associative relationships to provide a query expansion capability using SPARQL 1.1 property paths.

The case study shows that it is possible to semantically integrate information extracted from datasets and grey literature reports in different languages and provide KOS-based search. The presentation reflects on lessons learned, including the need to allow resources for extensive data cleansing. Although more work on the NLP extraction methods is needed for an operational capability, the study was able to generate CRM/AAT based RDF from English, Dutch and

Swedish texts in the same format as that derived from the datasets, thus allowing cross search. A pattern based mapping methodology helped ensure the validity and consistency of the ontology mappings and the lower level implementation details. The Demonstrator also illustrates the possibility of domain application oriented user interfaces for searching RDF datastores. Automatically generated metadata from natural language does not have the same reliability as metadata automatically derived from datasets (after data cleansing); future work should express the provenance of the subject metadata extracted and also the method by which it was extracted. Details of the case study methods and results can be found in Binding et al. (2018).

Acknowledgements

This work was supported by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193 (the ARIADNE project).

References

AAT. (2018). Getty Art & Architecture Thesaurus as Linked Open Data, Getty Vocabulary Program, Retrieved May 5, 2018, from http://vocab.getty.edu/

ARIADNE. (2017). ARIADNE Project. Retrieved May 5, 2018, from http://www.ariadne-infrastructure.eu

ARIADNE Portal (2017). Retrieved May 5, 2018, from http://portal.ariadne-infrastructure.eu/

Binding Ceri, Michael Charno, Stuart Jeffrey, Keith May and Douglas Tudhope. (2015). Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data. International Journal on Semantic Web and Information Systems, 11(1), 1-29.

Binding Ceri, Douglas Tudhope and Andreas Vlachidis. (2018). A study of semantic integration across archaeological data and reports in different languages, Journal of Information Science, Retrieved Aug 22, 2018, from https://doi.org/10.1177/0165551518789874. (an open access 'author accepted version' is available at https://pure.southwales.ac.uk/files/2683350/Archaeology_integration_JISauthorversion2.docx).

Demonstrator. (2017). Demonstrator for dendrochronological data integration case study. Retrieved May 5, 2018, from http://ariadne-lod.isti.cnr.it/description.html

Isaac Antoine, William Waites, Jeff Young J and Marcia Zeng. Eds. (2011). Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report, Retrieved May 5, 2018, from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/

STELETO. (2016). STELETO open source code, Retrieved May 5, 2018, from https://github.com/cbinding/steleto/

Tudhope Douglas, Keith May, Ceri Binding, Andreas Vlachidis. (2011). Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, Retrieved May 5, 2018, from https://doi.org/10.11141/ia.30.5

Designing a Multilingual Knowledge Graph as a Service for Cultural Heritage – Some Challenges and Solutions

Valentine Charles
Europeana Foundation,
The Netherlands
valentine.charles@europeana.eu

Hugo Manguinhas
Europeana Foundation,
The Netherlands
hugo.manguinhas@europeana.eu

Antoine Isaac
Europeana Foundation,
The Netherlands
antoine.isaac@europeana.eu

Nuno Freire
INESC-ID, Portugal
nuno.freire@tecnico.ulisboa.pt

Sergiu Gordea
[3]AIT Austrian Institute
of Technology, Austria
sergiu.gordea@ait.ac.at

Abstract

Europeana gives access to data from Galleries, Libraries, Archives & Museums across Europe. Semantic and multilingual diversity as well as the variable quality of our metadata make it difficult to create a digital library offering end-user services such as multilingual search. To palliate this, we are building an “Entity Collection”, a knowledge graph that holds data about entities (places, people, concepts and organizations) bringing context to the cultural heritage objects.

The diversity and heterogeneity of our metadata has encouraged us to re-use and combine third-party data instead of relying only on those contributed by our own providers. This raises however several design issues. This paper lists the most important of these and describes our choices for tackling them using Linked Data and Semantic Web approaches.

Keywords: linked data; knowledge graph; Europeana.

1. Introduction

Europeana gathers over 50 million paintings, books, newspapers, audio recordings, etc., from more than 35 European countries and in more than 40 languages. With such a diversity, supporting users in their (multilingual) search and browsing activities is a challenge. The vision of Linked Open Data applied in the cultural sector (Gradmann, 2010) has led us into collecting more data about contextual entities such as people, places, concepts next to Cultural Heritage Objects' (CHOs) metadata. The Europeana Data Model (EDM) (Europeana, 2016) enables our data partners to describe contextual entities as Linked Data (LD) resources with their own URI identifiers instead of literals. In addition, to increase the semantic and multilingual coverage of its metadata, we perform automatic semantic enrichment of our dataset by linking literals found in the CHO metadata to linked open multilingual datasets such as GeoNames[1] and DBpedia[2] - see documentation and examples at (Europeana, 2018). The number of links between CHOs and contextual entities as well as of data containing multilingual labels has thus grown considerably. However, this richer data is still heterogeneous: different providers use resources with different, not necessarily entirely commensurate, semantic and multilingual characteristics, while others do not use any such resources at all.

To palliate this, we have begun to select and combine statements from various LD sources into an "Entity Collection" (EC), a knowledge graph (KG) centralising data about contextual entities.

[1] http://www.geonames.org/

[2] http://wiki.dbpedia.org/

The EC is intended for use by several Europeana services, most immediately as a means to improve the users' experience in their search for CHOs (Hill et al., 2016a). It is designed to enhance:

- **Findability:** users can refine their search by filtering and browsing on people, places and subjects. Using the EC data for semantic enrichment reduces ambiguity in the CHO metadata, clarifying its meaning and improving its interlinking. Multilingual search benefits significantly from the multiple labels typically associated with each entity. For instance, an Entity auto-completion feature would use the EC to power search by keyword, returning a list of entities that have a label that matches what the use has typed, for any language available in the EC.
- **Contextualisation:** users can see additional contextual information related to specific CHOs. The EC can support annotation scenarios (semantic tagging) by suggesting entities to be used as tags instead as mere strings.
- **Exploration:** users can browse the relationships between CHO resources and entities. For instance, if an Entity created a CHO, a user could access the CHO via the page dedicated to that Entity, or access to more details about the Entity from the CHO item page.

The building of the EC has raised several challenges, motivating design decisions and solutions that we report in this paper. Section 2 presents related work on the activities involved in the creation, population, sharing and re-use of KGs. Building a KG such as the EC as an operational service requires well-designed processes for importing entities from external data sources and making the data available for exploitation, while maintaining data integrity and freshness as these sources evolve. The main activities and automatic processes involved are presented in Fig.1 and described in sections 3 and 4. We finish with a summary of our activities and future work.

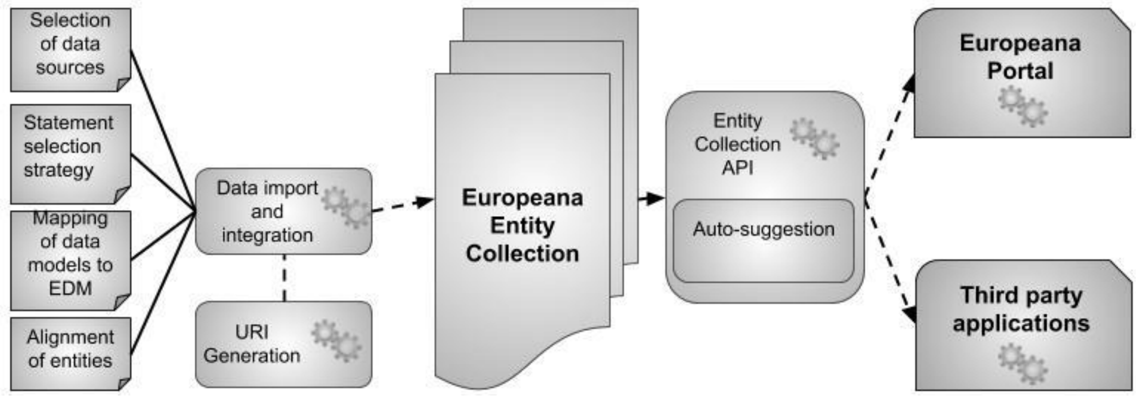

Selection of data sources
Statement selection strategy
Mapping of data models to EDM
Alignment of entities
Data import and integration
URI Generation
Europeana Entity Collection
Entity Collection API
Auto-suggestion
Europeana Portal
Third party applications

FIG. 1: Overview of Entity Collection processes in Europeana

2. Related Work

KGs have been created to solve data heterogeneity and quality issues, to structure and organise back-end datastores, and to provide advanced end-user services. They are typically intended to unify and enhance existing data, providing a centralised service capable of addressing issues of (query) disambiguation, responsiveness, relevance ranking, data enrichment, etc.

The best-known KG implementation is perhaps Google's Knowledge Graph, which exploits information extracted from a number of web sources (Dong et al., 2014). In the LD community, DBpedia has long played a key role for providing a large, open body of knowledge that others can re-use and link to (Auer et al., 2007). Wikidata[3] is another example of a (crowdsourced) open database, which is also used as a data source of Google's own KG.

[3] http://wikidata.org

Gabrilovich & Usunier (2016) presents the many research aspects involved in the creation of KGs: relation extraction, conversion and mapping, ontology matching, etc. Not all of these, however, are relevant for Europeana. For example, Dong et al. (2014) and Szekely et al. (2015) focus on the problems of knowledge extraction and merging from large set of automatically extracted data, including unstructured and structured sources. We do not aim to operate at such scale, instead focusing on building a KG on top of already extracted and structured knowledge.

DBpedia and Wikidata integrate different sources too. But their information-orientation is different. DBpedia extracts data from semi-structured sources in one information space (Wikipedia). Wikidata sources data from the crowd. In both cases there is no range of pre-existing external 'official' sources. In particular, the modelling of the data can be decided based on what is available (and needed) in the 'information ecosystem', which is directly at hand. There can be conflicts in the data though, i.e., statements reflecting views of different Wikidata contributors (or their sources). To address this, Wikidata handles provenance at a very granular level (individual statements). Multilingualism - a key issue for us - is also a focus in both initiatives: DBpedia separates language editions, but seeks to interconnect them as much as possible, while Wikidata starts with language-neutral resources and adds language-specific information about them.

BabelNet4 is another KG that heavily focuses on multilingualism. It links some 16 million entities across 284 languages. In terms of data integration, it sits 'above' Wikidata, including it as a dataset alongside many other data sources, including GeoNames and Wordnets for various languages (Navigli & Ponzetto, 2012). Like some other KGs, it is also not open enough: its license prevents the sort of partial re-publication Europeana performs to provide its (open) services.

Other relevant work includes efforts on tackling specific problems of KG creation. A lot of work in the domain focuses on ETL aspects, such as mapping and conversion of one dataset into a KG (Pellissier et al., 2016): but unlike many of these efforts, our EC is not about publishing legacy data as LD. Rather, we are re-users of already published and curated data. In addition, we do not need to represent all the information from the data sources that we re-use for our KG: we can and should focus on the most useful parts for us and our re-users5. We expect that designing our EC needs to combine automatic and manual processes where the organizational setting is clear and that it will in the first instance benefit from wider discussions on management of data flows such as versioning, archiving and on the documentation of changes, along the lines of the OAIS reference model ("Open Archival Information System ", n.d.).

More directly relevant to our case, considerable work has been devoted to "reconciliation" (aka. "matching" or "alignment") of entities across datasets. This is a vital concern for Europeana, as the sources we seek to use can have overlapping scopes. Automatic matching (Euzenat & Shvaiko, 2013) as well as manual and semi-automated approaches (Ossenbruggen et al., 2011) can be relevant here. The problem can be also mitigated by selecting sources (or parts thereof) with very limited (or no) overlap.

We envision our KG as being built by in-house specialists in cultural-sector data, and we count on our active network of data partners to flag relevant data sources to integrate, e.g., because their scope would match well their datasets. Instead, related work in search can be more relevant for our attempts to provide discovery services, especially searches for entities, ranked by their relevance, as e.g. Google provides for their KG (Google, 2018). (see Section 4.3 for our choices on ranking)

[4] http://babelnet.org/

[5] For example, the DBpedia to EDM mapping only captures the information Europeana needs: https://github.com/europeana/tools/blob/master/europeana-enrichment-framework/enrichment/enrichment-framework-knowledgebase/src/main/resources/dbpedia2agent.xsl

General best practices for publishing data are also relevant. The W3C recently published Data on the Web Best Practices (Farias Lóscio et al., 2017) with recommendations such as "reuse vocabularies, preferably standardized ones", which especially argues for not re-inventing the wheel in terms of the classes and properties used to express structured data. Europeana does not refrain from minting its own classes and properties when needed. But the position of our EC as a service built on top of existing data and which needs to remain interoperable with the data others publish in our community, raises a strong requirement for re-using existing ontologies. This is a difference with e.g. DBpedia and Wikidata, which create specific ontologies and align them afterwards with existing vocabularies when possible. Szekely et al. (2015) have adopted an existing ontology, Schema.org6, which is also used by Google. Another recommendation is to "make data available through an API". We aim to make available, at a minimum, an entity discovery service, alongside raw access to data via LD content negotiation for entities, batch dump access and an expert (and difficult to maintain) SPARQL endpoint. Like Google, DBpedia provides a simple text-based entity look-up service. Wikidata provides the full MediaWiki API, geared towards the retrieval of Wiki pages; access to data is chiefly handled through the LD content negotiation, dumps and a full SPARQL query service.

The sector of Galleries, Libraries, Archives & Museums (GLAM) has recognized early the potential of Linked Open Data and several efforts have been carried out, which can be compared to ours. Organizations have released contextual entities from their legacy vocabularies, gazetteers and authority lists. Concepts, person names and place names from the Getty Museum Art and Architecture Thesaurus (AAT), Union List of Artist Names (ULAN) and Thesaurus of Geographic Names (TGN) are available via content negotiation and a SPARQL endpoint (Getty, 2018). The German National Library has published its reference set of resources (GND) as LD (DnB, 2018a) similarly to the French, American and Spanish National Libraries.

While these efforts chiefly aim at publishing data from relatively isolated (institutional) information spaces, they try to create links to other datasets, starting with their peers. Some projects are dedicated to 'network' reference datasets. OCLC's Virtual International Authority File7 (VIAF) merges person and organization data from authority lists from more than 50 national libraries and agencies. It serves a unified description of each authority next to links and the original data from each library, see for example: http://viaf.org/viaf/9847974.rdf. The German National Library runs the Entity Facts service serving GND data combined with other datasets, including VIAF (DnB, 2018b). The SNAC project8 has performed a merging of data for persons found in archive collections. It connects its data to others, such as Getty's ULAN. Cross-datasets links can already be present in the original data or require semi-automatic reconciliation. Often a mixture of both happens, i.e., legacy identifiers from external datasets are found in the records of a source dataset and these implicit links need to be made explicit as URI references (e.g. https://www.europeana.eu/portal/en/record/90402/SK_A_4691.html which has identifiers from the Rijksmuseum and Europeana).. This renders the alignment processes often very specific to the data at hand - say, library and archive records could use quite different matching scripts.

The thematic project Europeana Food and Drinks has performed an interesting experiment, selecting relevant concepts from general datasets like DBpedia and linking them to institutional datasets to form a common "classification" for the project (Alexiev, 2015). They compared the multilingual interest of the various options available. This is similar to what we intend for our EC. We need to address a wider scope across subjects and types of collections, however, as well as publish our data in channels that can serve more purposes.

Note that despite their specificities we can benefit from these GLAM-related efforts from a data representation perspective, as most of them adhere to the principle of re-using existing ontologies. Some are also great examples regarding the distribution of the data. For example, the

[6] http://schema.org

[7] http://viaf.org/

[8] http://socialarchive.iath.virginia.edu/

STW thesaurus for economics has a web service9 that is exemplar of the way SKOS-like concept vocabularies can be served via a web API. DigitalNZ, a GLAM aggregator like Europeana, provides a Concepts API for its data re-users (DigitalNZ, 2015). Finally, OCLC's Worldcat Identities project (O'Reilly, 2007) is a good example of how entities can be used to provide novel ways to find and explore objects.

3. Building and Making Available a Knowledge Graph for Europeana

Europeana data experts and officers take the strategic decisions needed to import, integrate and manage data in our KG, including criteria to select data sources, and maintain our data model to represent and map the entities to the data. They perform the configuration and regularly execute the import and update of entities, which are then made available through a dedicated API.

3.1. Selection of Data Sources

Selecting data sources (or parts thereof) to integrate in the EC requires an intellectual effort prior to the actual harvesting and import of the data. It implies analysis of external data by a data expert and application of selection criteria. Europeana's strategy relies on leveraging existing linked open datasets and vocabularies and the following criteria to evaluate and select data sources (Isaac et al., 2015):

- **Availability and Access:** The datasets should be available on the Web and compliant with the LD recipes. They should be re-usable under an open license.
- **Granularity and Coverage:** The datasets should have the same coverage or should obviously complement each other. Reconciling resources that are semantically too far from each other could introduce ambiguities or semantic flaws for entities. For Europeana the data sources should answer to Who?', 'What?', 'When?', 'Where?' questions that are the most relevant to the cultural heritage domain as they help contextualise CHOs. Language coverage is also a key requirement: we aim to support over 29 languages in which Europeana receives metadata as reported in (Hill et al., 2016b)·. Ideally a dataset should provide labels in all the languages supported by Europeana or contribute with the labels necessary to reach such coverage. Generic data sources in terms of coverage or granularity are also likely to introduce semantic flaws during manual or automatic enrichment processes (see below on 'size').
- **Quality:** This includes intrinsic aspects of the dataset that can be manually or automatically assessed, such as the structure and representation of values and languages.
- **Connectivity:** The richness of the EC will be improved if the selected datasets have incoming and outgoing links to other datasets.
- **Size:** Depending on the size of the selected dataset, the number of entities is a criterion of selection. A high number of resources and statements is preferable, if the alignment process can deal with the greater ambiguity (i.e., higher number of entities associated with a given name) that larger sizes tend to generate. For example, GeoNames has 7.5M place names. The name "Guadalajara" limited to Mexico returns over 15 places, a lot of them are small *pueblas* with population under 15.

The need for a consistent and value-adding EC dictates a careful strategy for balancing domain-specific sources with more generic ones while addressing issues of semantic grain mismatch. We tend to choose general "pivot" datasets to cover as many entities as possible. For instance, Europeana might favour Wikidata over domain specific vocabularies such as Getty's AAT. Yet, in some cases we may want to give precedence to complementary datasets for more specific entities. Complementarity is not only relevant for entity-level data but also for CHO-level metadata: for instance, a dataset that includes metadata for CHOs could be used to create abstract "work"-level entities for our own CHOs, as it is often the case in library metadata. Note that the question of selecting pivot data sources vs. complementary (or domain) ones is

[9] http://zbw.eu/beta/econ-ws/about

independent from the actual alignment of entities in the EC (whether merging entity resources or representing matches between them as links, which preserves the original data).

The next step is to choose entities to be imported in the EC. The manual selection of individual entities from a data source is time-consuming and unfeasible for large sources. A query scenario is therefore envisioned, where a user can define the selection by designing queries to a data source (if a query service is available) that implement the appropriate selection criteria. For instance, in order to only import in the EC DBpedia data related to artists, a filter query would be created based on the statement pattern *anEntity rdf:type dbp:Artist* .

3.2. Data Modelling, Mapping and Statement Selection

Building a KG requires data to be represented in a consistent way. Each linked entity in the EC is an instance of a contextual class as defined in the EDM for representing people (*edm:Agent*), places (*edm:Place*), concepts (*skos:Concept*), time periods (*edm:Timespan*) or organizations (*foaf:Organization*). Mappings are created between the data model of a selected data source and EDM[10]. Custom mappings to EDM are needed to select the relevant information and the properties for given entities. This process is made easier (if not trivial) when the data sources are based on SKOS (Simple Knowledge Organisation System) (Miles & Bechhofer, 2009) which EDM re-uses for describing concepts and also preferred and alternatives labels for people, places, time periods and organizations. Note that besides the top-level classes above, most of the EDM elements[11] come from ontologies used in (cultural heritage) linked datasets, such as Dublin Core, RDA, and FOAF. EDM also seeks to adhere to the W3C best practice "choose the right formalization level": we refrain from adding too many formal axioms that would make mappings harder and perhaps disqualify good data sources without a serious reason besides elegance of modelling.

We also use mappings to select statements to be imported in the EC, e.g. by filtering out properties, (sub-)types of entities or specific resources (URIs), if they are irrelevant for Europeana. Note that Europeana does not need every statement from the selected datasets, e.g., labels for languages that it does not support (in GeoNames) or entities not relevant for Cultural Heritage such as modern pop stars (in DBpedia)[12].

3.4. Data integration, Reconciliation, Alignment and Curation

After being imported in the EC, the new entities need to be integrated with the existing EC entities. This step consists in the following workflow – some components of which have been already implemented as part of the semantic enrichment mentioned earlier:

Integration and reconciliation of entities. Imported entities are integrated with existing EC entities (i.e., the statements about these two entities are merged) or new corresponding entities are created (i.e., a new Europeana URI is minted). This is supported by the execution of automated background data-processing jobs, with scheduling, notification and reporting functionalities. Entity data will be previewed before integration into the EC for quality control purposes. The integration strategy may be influenced by the selected data sources. For instance, using Wikidata as a pivot data source for all the Europeana entities would make it easier to reconcile entities within the EC, as it is very rich in alignments to datasets in our sector (e.g., VIAF). Wikidata would then be used as a source from which Europeana could access other vocabulary alignments.

Alignment of entities. The detection of duplicates within the EC is currently based on the co-referencing information found in the data (*owl:sameAs* or *skos:exactMatch* links). We do not

[10] The mappings we use for the EC source datasets (DBpedia, GeoNames, etc) can be found at https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/enrichment/enrichment-framework-knowledgebase/src/main/resources

[11] See a full listing at https://github.com/europeana/corelib/wiki/EDMObjectTemplatesEuropeana

[12] See for example the list of filtered agents: https://docs.google.com/spreadsheets/d/1Wu8gPsgdtwnDN-GSuettT8WwqmvTeHaeAlqBF8-_joE

exclude the possibility of creating alignments using (semi-)automatic or manual tools such as Mix'n'match[13] and CultuurLink, following up on recent experiments (Manguinhas et al., 2016). We have found that despite selecting large datasets we are still missing a lot of coreferencing information to other datasets (chiefly domain vocabularies, but also reference datasets such as VIAF).

M*anual curation of entities and/or data*. As an additional step to maintain integrity, curators from Europeana staff will be able to edit the data for a Europeana entity by adding, changing or removing statements (including alignments), without preventing future updates from the imported data sources. Existing entities may also be deprecated.

These workflows will also benefit from additional normalisation and cleaning rules to apply to the data collected for each entity, as hinted from some "matching rules" presented in the documentation of Europeana's automatic semantic enrichment (Europeana, 2018). For instance, labels and values are not always accurate, and are sometimes even missing.

3.5 Data Integration Strategies

The management of the data within the EC has raised key data integration problems, which we are still discussing at the time of writing.

The main issue concerns when descriptions coming from different sources require merging, i.e. whenever two or more resource descriptions exist for the same entity. A choice is needed regarding which statements will be prioritised to become part of the description for the resulting Europeana entity. We have identified several options:

- **Unification**. The simplest strategy is to unify all statements coming from the different datasets into a single description. However, this strategy may lead to inconsistencies, e.g. cases where more than one statement exists for the same property when only one is allowed (e.g. the birthplace of a Person is stated in source A to be a country while source B is more granular and states the city) and contradictory statements (e.g. two distinct birth dates for the same Person).
- **First come / first serve**. This strategy considers an order (for the source datasets) while selecting the statements for the entity description. While copying a statement in the EC, the cardinality constraints defined for a given property are enforced by skipping the statement once the maximum is reached. The order in which the source datasets are merged may be defined to reflect the distinction between the pivot and complementary datasets, so that a pivot takes precedence by being the first to be considered for merging.
- **Most representative**. This strategy chooses among conflicting statements based on the number of source datasets that contain them. This assumes that if a statement is found in more datasets, it is more likely to be "true". However, there can be situations where incorrect statements may be spread, as many datasets integrate data from other sources, replicating the issue. Also, the strategy does not define how a statement can be chosen in case of a tie.
- **Differentiated most representative**. This more complex strategy tries to balance pros and cons from the previous strategies by distinguishing the datasets into two explicit groups (pivot and complementary). For competing statements *within* a group, this strategy may apply the "most representative" or the "first come / first serve" strategies. Then, statements from the pivot group are copied, and statements from the second group are added - while preserving cardinality constraints.

Any chosen integration strategy will be supported by provenance and attribution information capturing the source of a given entity or statement (e.g. tracking the source URIs in an *owl:sameAs* or *skos:exactMatch* for an entity or RDF Graphs for statements).

[13] http://tools.wmflabs.org/mix-n-match/

3.6. Data in the Entity Collection

The current data available in the EC inherits from the data sources previously harvested to underpin Europeana semantic enrichment. As of May 2018, the EC contains data for:

- 215.802 Places: a subset of **Geonames**, corresponding to places part of European countries and of a specific feature class[14]. (i.e. "A", "P.PPL", "S.CSTL", "S.ANS", "S.MNMT"...)
- 165.005 Agents: a subset of **DBpedia** corresponding to most of the instances of *dbp:Artist* with some exceptions, and integrated from 49 DBpedia language editions. All locale DBpedias that match the list of languages supported by Europeana have been harvested from which a selection is made to enrich concepts and persons.
- 1.572 Concepts: a subset of **DBpedia** comprising a handful of WWI battles, the "World War I" category and other categories[15] being used for Europeana Collections. And also two vocabularies: one for music genres, forms and compositions obtained from Wikidata and the photography vocabulary maintained by the Photo Consortium.
- 599 Organizations: data about Europeana's data partners collected through our Customer Relationship Management (CRM) system. Co-references to Wikidata were added when available and represented as *owl:sameAs* relations.

We will add more entities, first from the data sources we already ingest, and then extending to other data sources, especially Wikidata (see Section 3.1 for our motivations), as well as time spans, which are not yet represented in the EC.

4. Accessing the Entity Collection Data

The EC is made available via an API ("Europeana Entity API", n.d.), which powers the search query auto-completion and the entity pages in Europeana.

4.1 Entity Collection Look-up API

Two API methods are available to look up for entities in the EC. The first one uses content negotiation to deliver data in HTML or JSON-LD formats, according to the client preferences indicated through the HTTP request header. A known entity can be accessed using its URI; the content negotiation service will automatically redirect the request either to the Entity API endpoint, or to the Entity Page in Europeana16.

The second method enables to look up an EC entity using an alternative URI that is recorded in the source dataset. This lookup uses the owl:sameAs and skos:exactMatch co-reference statements available within the entity data and returns a redirection in line with common HTTP best practices. This method is a key requirement for semantic integration of Europeana KG with the existing linked data repositories.

The default format chosen for representing the entities and facilitate the re-use of the data in the EC is JSON-LD (Sporny at al. 2014), the JSON representation for LD. This format was chosen as it is commonly used in Web-based programming environments, to build interoperable Web services. It can also be used when data is integrated in other pieces of JSON data, such as the ones returned by the autosuggestion API (see Section 4.3). To make the JSON-LD serialisation more compact, we have defined a JSON-LD context, which defines abbreviations for the namespaces used in EDM and specific data types (e.g., http://rdvocab.info/ElementsGr2/gender can be simply referred to as "gender"). The data thus becomes better understandable by (third party) web developers without affecting the underlying semantics. Some EDM properties can be used with several values in different languages, such as

[14] http://www.geonames.org/statistics/total.html

[15] See: https://docs.google.com/spreadsheets/d/1qjyyneg6aMoPC2v5hwC8YinmHKNyJtvTJp1HJdnnPc8

[16] For instance, http://entity.europeana.eu/entity/agent/base/146741?wskey=apidemo. NB: at the time of writing one still needs a key to de-reference these URIs. This will be changed later.

skos:prefLabel, skos:altLabel, foaf:name. To facilitate standardisation, this context is available as a separate resource17 which can be referenced in the JSON-LD serialization of the contextual entity.

RDF/XML will be also supported as it is commonly used, especially for CHO metadata ingestion at Europeana.

4.2 Generation of URIs in the EC

An important aspect of data integration in the EC is the generation of URIs for every entity. Our design is based on a LD scenario where URIs must be (i) **Dereferenceable**, both humans and user-agents must be able to meaningfully resolve the URI (ii) **Unambiguous**, a URI should not refer to two distinct resources (iii) **Immutable**, it should not change in time. As Europeana holds data which is not available elsewhere as a whole, it needs to create URIs in its own namespace (data.europeana.eu), so that a data consumer can access and retrieve the data. Identifiers need to be both easy to assign and future-proof. URIs follow the pattern: *http://data.europeana.eu/{entity_class}/{scheme}/{localID}*

- {entity_class} corresponds to the types of EDM contextual entities (Agent, Place, Concept and Organizations).
- {scheme} represents a sub-division under each entity class. A special division with the name "base" will contain all entities that are integrated from external data sources.
- {local_id} is the local identifier for the entity.

For the local identifier we chose to generate a sequential identifier for entities that are collected from external sources since it is the type that requires less effort to assign and maintain (Archer et al., 2012). The choice of minting human readable URIs was discussed and rejected within our community (Europeana, 2015) as it increases complexity for both maintenance and data consumption. Such URIs could be envisioned as alternative URIs. A more practical alternative to human readable URIs is to have URLs that, after content negotiation, contain a human readable part. This would have no impact on data consumption and would require considerably less effort to implement and maintain.

4.3 Discovery of Entities

The API provides another two methods for discovery and retrieval of entities in the EC:

- *entity auto-completion:* implementing quick search by entity names. This type of discovery, integrated in Europeana to support end-users to formulate more precise search queries, is based on entity labels only (i.e. *skos:prefLabel, skos:altLabel, edm:acronym*).
- *entity search:* supporting retrieval of entities by using free querying on all properties or on (a combination of) individual properties. The latter enables advanced search scenarios, e.g. finding cities in a given country (using *edm:isPartOf*), or fashion designers born in the XIXth century (e.g by using *rdagr2:professionOrOccupation* and *rdagr2:dateOfBirth*)

Recommending entities for search auto-completion is a challenge, given the requirement for achieving a high precision for suggestions in the top 10 list. Moreover, the multilinguality of the EC and the search queries (users often search in Europeana using their native language) add to the difficulty. The ranking of individual entities uses a formula that integrates and normalizes two measures: relevance and popularity. The relevance of an entity is computed as the number of Europeana records that contain one of the entity labels, while its popularity is computed using the Wikidata PageRank, as calculated across 133 of its languages versions (Diefenbach & Thalhammer, 2018). Preliminary testing has indicated that this approach yields good results, though the need for cross-linguistic matching due to the modest average multilingual coverage

[17] http://www.europeana.eu/schemas/context/entity.jsonld

currently limits the precision of the suggestions. Future work will investigate the employment of a Learning-To-Rank approach to improve the ranking of individual entities based on the information captured within the Europeana access logs.

The entity search has a generic implementation, allowing API users to formulate complex queries following the Solr query syntax ("SOLR Query Syntax," n.d.). Built-in statistics on EC are made available via facet profiles. For example, the faceted field on the property type provides, in real-time, the number of Agents, Concepts, Places and Organizations available in the EC (see also Section 3.6). The presentation of the search results uses pagination as specified by the Linked Data Protocol (Speicher et al., 2015). Applications that integrate search can thus easily fetch all results by issuing a chain of calls for the *next* (page) URL, which is available in every response.

5. Conclusion and Future Work

This paper has presented different requirements, highlights challenges and proposes solutions to adopt when building a knowledge graph for cultural heritage.

Solutions to some of the problems and questions raised in the paper have been found sufficient to allow the creation of a first version of the EC. However, some decisions still need to be taken to ensure the coherence of the EC over time

Data coverage and Extensibility. Europeana needs to expand its EC to cover as many CHOs as possible and support 'client' Europeana services. Future work includes the sourcing of suitable datasets to represent times periods as well as named events.

Data integration strategy. Both automatic and manual curation approaches need to be considered. Future work includes the improvement of the quality of current data by removing statements with no or faulty language tags, filtering unwanted statements or entities, refining the data mappings to include new statements, etc.

Enrichment. The EC will be used to enrich the Europeana metadata still represented as literals (the process mentioned above still uses a separate database).

Discoverability. The mapping work from Schema.org to EDM (Wallis et al., 2017) will allow the entities to be indexed by search engines and therefore more discoverable for the users. For instance the inclusion of owl:sameAs links from the Google Knowledge Graph in Schema.org markup would maximise the chance of the Europeana content to be displayed in KG cards.

References

V. Alexiev. (2015). Europeana Food and Drinks Deliverable D2.2 Classification Scheme. http://foodanddrinkeurope.eu/wp-content/uploads/2015/12/D2.2-Classification-scheme.pdf

P. Archer, S. Goedertier, N. Loutas. (2012). D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. Project Interoperability Solutions for European Public Administrations. https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives. (2007). DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web. Lecture Notes in Computer Science, vol 4825. Springer, Berlin, Heidelberg.

D. Diefenbach and A. Thalhammer. (2018). PageRank and Generic Entity Summarization for RDF Knowledge Bases. In: Lecture Notes in Computer Science, vol 10843. Springer, Berlin, Heidelberg.

DigitalNZ. Introducing the DigitalNZ Concepts API (2015). Retrieved August 13, 2018 from https://digitalnz.org/blog/posts/introducing-the-digitalnz-concepts-api

DnB - Deutsche National Bibliothek. (2018a). Linked Data Service of the German National Library. Retrieved August 13, 2018 from http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

DnB - Deutsche National Bibliothek. (2018b). Entity Facts. Retrieved from http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/entityFacts.html

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun,and W. Zhang. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Proc. 20th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining (KDD). 24-27 August, 2014, New York, USA. 601-610.

Europeana Foundation. (2015). Staying persistent: EuropeanaTech community help to pave the way for new Europeana URIs. Retrieved August 13, 2018 from http://pro.europeana.eu/blogpost/staying-persistent-europeanatech-community-help-to-pave-the-way

Europeana Foundation. (2016). Definition of the Europeana Data Model elements v5.2.7. Retrieved August 13, 2018 from http://pro.europeana.eu/edm-documentation

Europeana Foundation. (2018). Automatic Semantic Enrichment at Europeana. Retrieved August 13, 2018 from https://pro.europeana.eu/page/europeana-semantic-enrichment#automatic-semantic-enrichment

(n.d.). Europeana Entity API (alpha). In: Europeana Pro. Retrieved August 13, 2018 from https://pro.europeana.eu/resources/apis/entity

J. Euzenat and P. Shvaiko. (2013). Ontology Matching. 2nd Edition, Springer, 2013.

B. Farias Lóscio, C. Burle, N. Calegari (eds). (2014). Data on the Web Best Practices. W3C Recommendation. 31 January 2017. https://www.w3.org/TR/dwbp/

Getty Research Institute. (2018). Getty Vocabularies as Linked Open Data. Retrieved August 13, 2018 from http://www.getty.edu/research/tools/vocabularies/lod/

Google. (2018). Google Knowledge Graph Search API. Retrieved August 13, 2018 from https://developers.google.com/knowledge-graph/

E. Gabrilovich and N. Usunier. (2016). Constructing and Mining Web-scale Knowledge Graphs. Tutorial. In Proc. 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 17-21 July, 2016

S. Gradmann. (2010). Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. Europeana Whitepaper. Retrieved August 13, 2018 from http://pro.europeana.eu/publication/knowledgeinformation-in-context

T. Hill, D. Haskiya, A. Isaac, H. Manguinhas, and V. Charles. (2016a). Europeana Search Strategy. Europeana Whitepaper. Retrieved August 13, 2018 from: http://pro.europeana.eu/publication/europeana-search-strategy

T. Hill, A. Isaac, V. Charles, N. Freire and H. Manguinhas. (2016b). Europeana Search Strategy. Europeana Whitepaper. Retrieved August 13, 2018 from: http://pro.europeana.eu/publication/europeana-search-strategy

A. Isaac, H. Manguinhas, V. Charles, J. Stiller (eds). (2015). Selecting target datasets for semantic enrichment. Companion document to the report of the EuropeanaTech Task Force on Enrichemnt and Evaluation. Retrieved August 13, 2018 from https://pro.europeana.eu/project/evaluation-and-enrichments

H. Manguinhas, V. Charles, A. Isaac, T. Miles, A. Lima, A. Néroulidis, V. Ginouvès, D. Atsidis, M. Hildebrand, M. Brinkerink, S. Gordea. (2016). Linking subject labels in Cultural Heritage Metadata to MIMO vocabulary using CultuurLink. In: Proc. 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016) co-located with the 20th International Conference on Theory and Practice of Digital Libraries 2016 (TPDL 2016). Hannover, Germany, 9 September (2016).

A. Miles, S. Bechhofer, (eds.). (2009). SKOS Simple Knowledge Organization System – Reference. W3C Recommendation. https://www.w3.org/TR/skos-reference/

R. Navigli and S. Ponzetto. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, 217-250.

(n.d.). Open Archival Information System. In Wikipedia. Retrieved August 13, 2018 from https://en.wikipedia.org/wiki/Open_Archival_Information_System

T. O'Reilly. (2007). WorldCat Identities. In: O'Reilly Radar. [Blog post]. Retrieved August 13, 2018 from http://radar.oreilly.com/2007/02/worldcat-identities.html

J. Ossenbruggen, M. Hildebrand, V. de Boer. (2011). Interactive vocabulary alignment. In: Proc. 15th International Conference on Theory and Practice of Digital Libraries, Berlin, Germany, 26-28 September, 2011, 296-307.

T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher. (2016). From Freebase to Wikidata - The Great Migration. In Proceedings of the 25th International Conference on World Wide Web (WWW 2016), 11-15 April, 2016, Montreal, Canada, 1419-1428.

(n.d.). Solr Query Syntax. In Solr Wiki. Retrieved August 13, 2018 from https://wiki.apache.org/solr/SolrQuerySyntax

S. Speicher, J. Arwe, A. Malhotra (eds.). (2015). Linked Data Platform 1.0. W3C Recommendation. https://www.w3.org/TR/ldp/

M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, N. Lindström. (2014). JSON-LD 1.0, A JSON-based Serialization for Linked Data. W3C Recommendation. https://www.w3.org/TR/json-ld/

P. Szekely, C. Knoblock, J. Slepicka, C. Yin, A. Philpot, A. Singh, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, D. Stallard, S. Karunamoorthy, R. Bojanapalli, S. Minton, B. Amanatullah, T. Hughes, M. Tamayo, D. Flynt, R. Artiss, S. Chang, T. Chen, G. Hiebel, and L. Ferreira. (2015). Using a knowledge graph to combat human trafficking. In: The Semantic Web - ISWC 2015. Lecture Notes in Computer Science, vol 9367. Springer, Cham.

R. Wallis, A. Isaac, V. Charles, and H. Manguinhas. (2017). Recommendations for the application of Schema.org to aggregated Cultural Heritage metadata to increase relevance and visibility to search engines: the case of Europeana. Code4Lib Journal, 37 (April 2017). Available at http://journal.code4lib.org/articles/12330.

SESSION 3:
Metadata

National Diet Library Data for Open Knowledge and Community Empowerment

Saho Yasumatsu & Tomoko Okuda

Metadata as Content: Navigating the Intersection of Repositories, Documentation, and Legacy Futures

Erik Radio

Wikidata & Scholia for scholarly profiles: the IU Lilly Family School of Philanthropy pilot project

Mairelys Lemus-Rojas & Jere Odell

National Diet Library Data for Open Knowledge and Community Empowerment
Presentation

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

Tomoko Okuda
National Diet Library, Japan
t-okuda@ndl.go.jp

Keywords: The National Diet Library, national library, Japanese library, Open Data, civic tech

Abstract

In 2008, the National Diet Library (NDL) established its “Policy of providing databases created by the National Diet Library” with the intention that the data created and provided on the Internet by the NDL would be easily used by third parties. The policy states that any third party can freely search and view the content, copy search results or, provided it is for non-profit purposes, use an API or other means to acquire NDL data automatically. Since then, the NDL has published APIs for many of its major systems, including NDL Search, Web NDL Authorities, and the Great East Japan Earthquake Archive. In keeping with the international trend toward the creation of a Semantic Web, these systems publish bibliographic and authority data as Linked Data in RDF/XML, RDF/Turtle, and JSON-LD formats. This means that the URI to an individual data record will remain unchanged irrespective of changes to server systems, thereby providing persistence for third-party applications. In other words, Linked Data from the NDL is carefully designed to be compliant with a wide variety of implementations on the web by third parties.

Usage of this Linked Data did not exhibit significant growth, however, until the NDL took specific steps to promote its use. We conducted a series of interviews with citizen coders and developers and identified two problems. First, the terms of use for this data are not open enough to encourage diverse reuse. Second, since individual coders were generally unfamiliar with bibliographic and authority data, the provider of this data was required to undertake assiduous outreach activities. The NDL addressed these challenges within the framework of its “Policy of providing databases created by the National Diet Library” by providing bulk download of open datasets that can be used without restriction for either profit or non-profit purposes as well as by taking part in public events related to open data and civic technology, which provided increased opportunities for introducing NDL data in communities throughout Japan. To this end, the NDL has partnered with two of the major open data competitions in Japan: the Linked Open Data Challenge Japan and Urban Data Challenge. The NDL also began to organize ideathons and hackathons to promote its data and services. In addition to these hands-on events, the NDL also hosted a low-key lecture series in 2016 and 2017, called The NDL Digital Library Café, which was open to anyone with an interest in this subject, regardless of the level of their ICT skills.

These outreach activities resulted many interesting and potentially useful initiatives, such as Linked Web NDL Authorities, Visualization of Publishing Trends in Japan from 1950 to 2017, How Active are Our Representatives?, and A Map of 19th Century Villages in the Greater Tokyo Area.

This presentation is a follow-up of our poster on Linked Open Data presented at the DCMI annual conference in 2015. In this year's presentation, I would like to demonstrate the NDL's efforts and achievements during the past three years in promoting the use of our data, while showcasing some of the best civic-driven applications and visualizations of library data.

References

National Diet Library. Use and Connect: How to link to NDL LOD. Retrieved August 10, 2018, from http://www.ndl.go.jp/en/dlib/standards/lod/usecase.html

Metadata as Content: Navigating the Intersection of Repositories, Documentation, and Legacy Futures
Presentation

Erik Radio
University of Arizona, USA
radio@arizona.edu

Keywords: metadata; repositories; documentation

Abstract

Documentary Relations of the Southwest (DRSW) is a dataset of bibliographic metadata derived from over 1500 reels of microfilmed documents that trace the history of the southwest from the 16th century until Mexico's independence in 1821. Originally made available to scholars through a now defunct proprietary repository, DRSW’s future is currently being assessed in the context of other repository solutions. While migrating content is a familiar scenario, this migration highlights key challenges in navigating the intersection of legacy design and possible futures for metadata curation and repository selection. This presentation deals with challenges revolving around three paradigms: metadata as content, system documentation generation, and metadata futures for indexing and integration.

In the repository, contextual metadata is commonly considered distinct from the content it describes. DRSW is an uncommon case, as none of the documents have been digitized; the metadata is the content. This presents unique issues since the original metadata creation for DRSW was not created under the guidance of a metadata professional and contains errors (e.g. typos, term inconsistency). As a result structure for measuring semantic loss in metadata was devised as way of preventing similar scenarios in the future, and will be discussed in this presentation.

The second paradigm revolves around the generation of documentation by systems. The selection of a system has significant effects on how metadata is processed, edited, and exported. However, while metadata can travel between systems, documentation does not always travel along with it. In addition to contextual documentation, it appears to be increasingly critical that there be system generated metadata. While some may object this exists in change logs and similar tracking files, these are not necessarily generatable for digital collection managers in a streamlined way. Further, they do not commonly include decisions as to why certain changes were made, or similar such decisions, which are critical for understanding the provenance of metadata. As a tool for mitigating semantic loss, further examination into system effects on metadata, or ‘Processual Documentation’, and what such a mechanism would entail will be discussed, particularly as it impacts DRSW.

The final paradigm to be explored regards what this particular scenario suggests for the future of metadata migrations and its use in repositories. As mentioned DRSW was created using a local schema. Because of the lack of documentation, local schemas such as this are not a sustainable option for migrations as multiple standards in a repository can lead to indexing troubles as well as possibly being confusing for users unless there is significant work internally towards ontology alignment. It was initially suggested that migrating the metadata to Metadata Object Description Scheme (MODS) would allow for the dataset to be more extensible into the future, but curators decided that it must adhere to the original, local standard. This provokes serious questions as to the ability of metadata to be integrated into a linked data environment. If it is critical for metadata to be preserved in its nascent form, then there must be further capacity for metadata to grown synoptically. Whether this is something that can be afforded by repository software (i.e. multiple views of an object), or through linked data projects is a topic that remains to be further discussed.

This presentation will outline the roadmap for DRSW as it fits into this trajectory as a way of facilitating a discussion following the presentation on how similar collections are being approached.

Wikidata & Scholia for scholarly profiles: the IU Lilly Family School of Philanthropy pilot project

Presentation

Mairelys Lemus-Rojas
Indiana University-Purdue University Indianapolis, USA
mlemusro@iupui.edu

Jere Ode
Indiana University-Purdue University Indianapolis, USA
jdodell@iupui.edu

Keywords: Wikidata, Scholia

Abstract

During recent years, cultural heritage institutions have become increasingly interested in participating in open knowledge projects. The most commonly known of these projects is Wikipedia, the online encyclopedia. Libraries and archives in particular, are also showing an interest in contributing their data to Wikidata, the newest project of the Wikimedia Foundation. Wikidata, a sister project to Wikipedia, is a free knowledge base where structured linked data is stored. It aims to be the data hub for all Wikimedia projects. The Wiki community has developed numerous tools and web-based applications to facilitate the contribution of content to Wikidata and to display the data in more meaningful ways. One such web-based application is Scholia which was created to provide users with complete scholarly profiles by making live SPARQL queries to Wikidata and displaying the information in an appealing and effective manner. Scholia provides a comprehensive sketch of the author's scholarship. This presentation will demonstrate our efforts to contribute data related to our faculty members to Wikidata and will provide a demo of Scholia's functionalities.

At IUPUI (Indiana University-Purdue University Indianapolis) University Library, we conducted a pilot project where we selected the 19 faculty members identified as core faculty from the IU Lilly Family School of Philanthropy to be included in Wikidata. The School of Philanthropy, located on the IUPUI campus, is the leading school in the subject in the United States. The scholarship produced by its faculty is known to be widely used. The goal of this pilot was not only to provide a presence in Wikidata for our faculty, but also for their publications and co-authors. As a result, we created 110 items to represent some of the works produced by the faculty members and 58 items for all co-authors. Moreover, we selected three publications and worked through their lists of references to contribute 39 cited publication items. Doing the additional work of adding co-authors and cited publications allowed us to start interconnecting works. For the creation of Wikidata items, we used a combination of semi-automated and manual processes. Making use of existing tools such as Source MetaData, QuickStatements, and Resolve Authors alleviated the manual labor, and allowed us to make contributions more efficiently. Once the items were created in Wikidata, we used Scholia to generate the scholarly profiles.

By building on existing bibliographic and metadata skills, academic libraries have the capacity to create and curate data about scholars affiliated with their institutions. Our pilot project is just a first step toward more efficient and systematic library-based contributions to Wikidata. We expect that the data sets we build in Wikidata will help our institution better understand and describe the value of its scholarly work in the study of philanthropic giving, nonprofit management, and all other research domains that are a core feature of our campus. In addition to providing value to our local institution, our contributions to Wikidata serve to build an open data platform maintained by the commons. Wikidata provides a welcome, open source alternative to share scholarly and

bibliographic data in a marketplace where publishers and other information companies work to capture this data and to profit from selling it back to universities.

SESSION 4
Categorisation

Why Build Custom Categorizers Using Boolean Queries Instead of Machine Learning? Robert Wood Johnson Foundation Case Study

Joseph Busch & Vivian Bliss

Categorization Ethics: Questions about Lying, Moral Truth, Privacy and Big Data

Joseph Busch

Why Build Custom Categorizers Using Boolean Queries Instead of Machine Learning? Robert Wood Johnson Foundation Case Study

Presentation

Joseph A Busch
Taxonomy Strategies, USA
jbusch@taxonomystrategies.com

Vivian Bliss
Taxonomy Strategies, USA
vbliss@taxonomystrategies.com

Keywords: automated categorization; Boolean query categorization; auto-classification; text analytics; recall and precision; Robert Wood Johnson Foundation

Abstract

This abstract provides an update on a project to build a Boolean query categorizer against a set of pre-defined broad categories for the Robert Wood Johnson Foundation (RWJF) a philanthropy dedicated to impacting health and health policy in the United States. Lessons learned building out the categorizer to make it scalable and maintainable are discussed.

1. Pre-defined Boolean Queries

In machine learning, all you need to provide is lots of content. The system figures out what it's about. But the problem with machine learning is that it is opaque, it's difficult to understand why an item is considered relevant. Categories are generic, may be irrelevant, can be biased, and are difficult to change or tune.

What if you want to categorize a collection against a set of pre-defined categories? One way to do this is to develop a set of Boolean queries that scope the context for each category. This is much more transparent than machine learning, and it provides relevant categories. But it requires a lot of work to set up, and specialized skills.

A Boolean query is a type of search that combines keywords or phrases with AND, OR, and NOT operators.

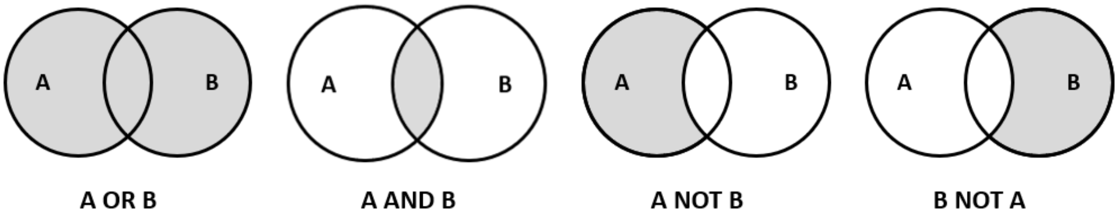


FIG. 1. Boolean query types illustrated using Venn diagrams.

Boolean queries are often used with proximity search. Proximity searching is a way to search for two or more words that occur within a certain number of words from each other, or within a section of a document. Unfortunately, Proximity operators and syntax are not standardized. The query syntax for Boolean queries also includes bounded phrases usually with quotations; right, left, and internal truncation; and nested statements with parentheses that match up.

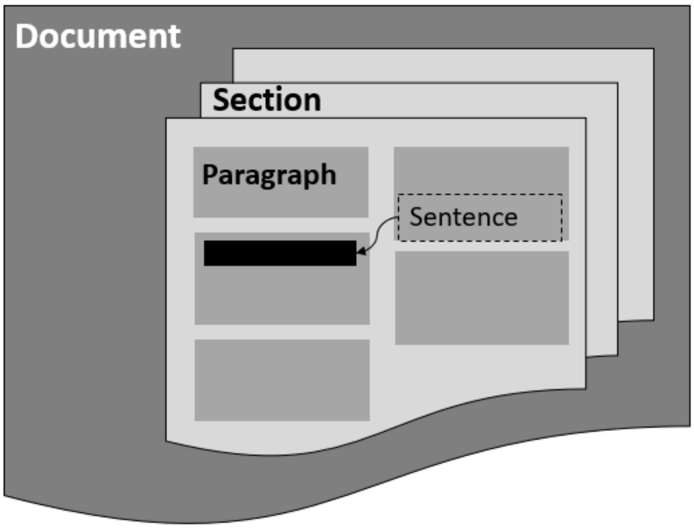


FIG. 2. Proximity searching specifies where query terms are located in documents.

2. Case Study

The Robert Wood Johnson Foundation (RWJF) is the largest philanthropy dedicated solely to health in the United States. Taxonomy Strategies has been working with RWJF to develop an enterprise metadata framework and taxonomy to support needs across areas including program management, research and evaluation, communications, finance, etc. We have also been working with RWJF on methods to apply automation to support taxonomy development and implementation within their various information management applications.

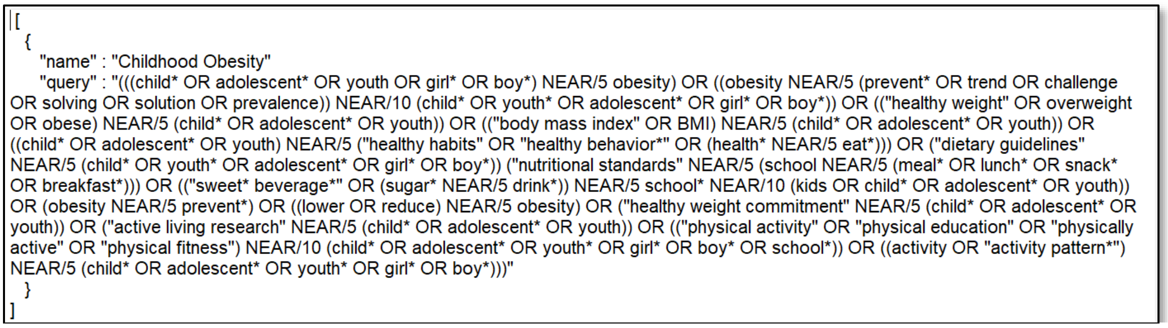
The initial target application for automated categorization is RWJF grant "precis" which are short descriptions of funded projects. Over the last five years, RWJF has made awards ranging from $3,000 to $23 million with time periods ranging from one month to five years. However, most grants are in the $100,000 to $300,000 range, and run from one to three years. (RWJF, 2018) RWJF grants are currently described with metadata including: Program Areas, Types of Support, Grantmaking Interventions, Demographics, Topics and Tags. But the existing descriptive metadata are difficult to use to accurately answer questions about grantmaking trends, thus staff do not use it. Taxonomy Strategies is working on a new metadata scheme and taxonomy to replace the current descriptive metadata. Automated methods will be critical for updating descriptive metadata from the current to the new metadata scheme and values.

In 2017, Taxonomy Strategies developed a pilot categorizer for 4 pre-defined Topics that describe some of the focus areas for RWJF programs and grantmaking – Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, and Health Coverage – using Lexalytics Semantria. (Lexalytics, 2018) This case study was presented in a DCMI Webinar on July 19, 2018. (Busch, 2018)

In 2018, Taxonomy Strategies is working with RWJF to: (1) develop requirements for, and suggest how to integrate text analytics and information retrieval software into RWJF staff workflows; (2) develop requirements for, and suggest how to build test collections for refining recall and precision for auto-classification; and (3) develop recommendations for staff roles and processes to support categorization of legacy assets and incoming grantee products.

1.1. Breaking down broad topics into simple queries

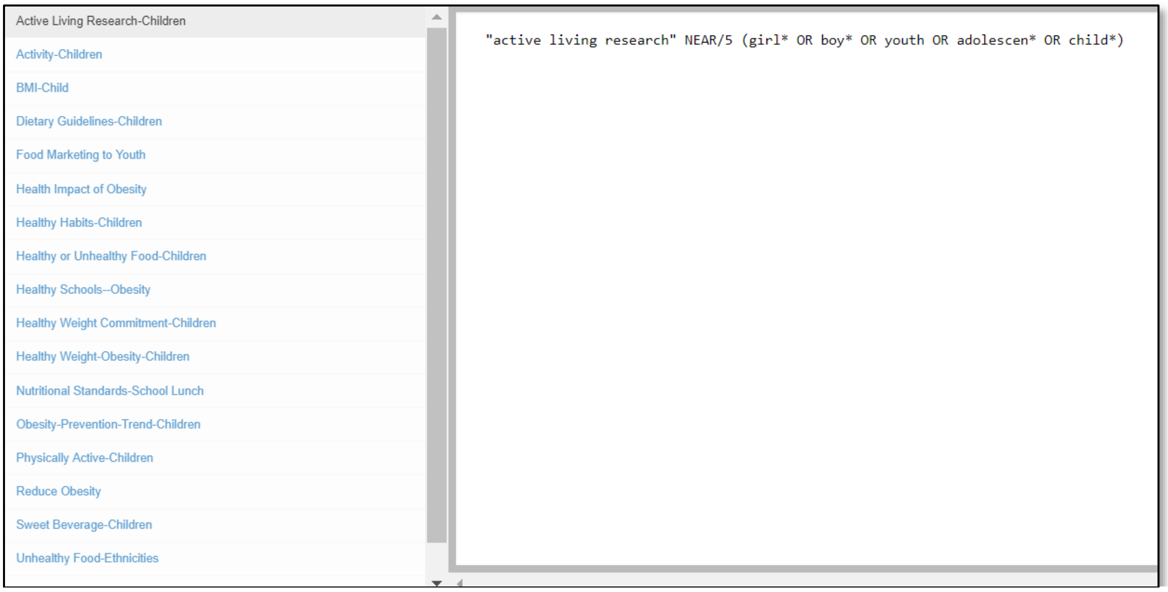
In the pilot project, Taxonomy Strategies built-up Boolean queries for the four target RWJF Topics. This was done using a text editor as shown in FIG. 3, then the complex query was cut and pasted into the Semantria Web user interface. Semantria validated the queries' syntax and either successfully loaded them or returned error messages which needed to be resolved. Eventually each of the four queries was successfully loaded.



```
[
  {
    "name" : "Childhood Obesity"
    "query" : "(((child* OR adolescent* OR youth OR girl* OR boy*) NEAR/5 obesity) OR ((obesity NEAR/5 (prevent* OR trend OR challenge
OR solving OR solution OR prevalence)) NEAR/10 (child* OR youth* OR adolescent* OR girl* OR boy*)) OR (("healthy weight" OR overweight
OR obese) NEAR/5 (child* OR adolescent* OR youth)) OR (("body mass index" OR BMI) NEAR/5 (child* OR adolescent* OR youth)) OR
((child* OR adolescent* OR youth) NEAR/5 ("healthy habits" OR "healthy behavior*" OR (health* NEAR/5 eat*))) OR ("dietary guidelines"
NEAR/5 (child* OR youth* OR adolescent* OR girl* OR boy*)) ("nutritional standards" NEAR/5 (school NEAR/5 (meal* OR lunch* OR snack*
OR breakfast*))) OR (("sweet* beverage*" OR (sugar* NEAR/5 drink*)) NEAR/5 school* NEAR/10 (kids OR child* OR adolescent* OR youth))
OR (obesity NEAR/5 prevent*) OR ((lower OR reduce) NEAR/5 obesity) OR ("healthy weight commitment" NEAR/5 (child* OR adolescent* OR
youth)) OR ("active living research" NEAR/5 (child* OR adolescent* OR youth)) OR (("physical activity" OR "physical education" OR "physically
active" OR "physical fitness") NEAR/10 (child* OR adolescent* OR youth* OR girl* OR boy* OR school*)) OR ((activity OR "activity pattern*")
NEAR/5 (child* OR adolescent* OR youth* OR girl* OR boy*)))"
  }
]
```

FIG. 3. Broad topic Boolean query from 2017 pilot project

In 2018, the process was modified to break up the broad topics into sets of simple queries. The goal was to make the queries more transparent, easier to "read", and easier to maintain as shown in FIG. 4. By "factoring" broad topics in constituent contextual parts, the simple queries could be combined and reused in different contexts. Working with simple contextual queries also facilitated "tuning" to optimize recall and precision.



Active Living Research-Children
Activity-Children
BMI-Child
Dietary Guidelines-Children
Food Marketing to Youth
Health Impact of Obesity
Healthy Habits-Children
Healthy or Unhealthy Food-Children
Healthy Schools--Obesity
Healthy Weight Commitment-Children
Healthy Weight-Obesity-Children
Nutritional Standards-School Lunch
Obesity-Prevention-Trend-Children
Physically Active-Children
Reduce Obesity
Sweet Beverage-Children
Unhealthy Food-Ethnicities

```
"active living research" NEAR/5 (girl* OR boy* OR youth OR adolescen* OR child*)
```

FIG. 4. Broad topic Boolean query broken up into simple queries.

1.2. Content collections for query building and testing

Choosing the content collection is a very important step in query building and testing. Busch (1998) suggests a "snowball" method to build up a collection starting with a list of relevant words and phrases to identify a core set of relevant articles from authoritative sources. Then performing a rhetorical analysis of titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) of the content items to build up a list of words and phrases and named entities. Iterating this process a few times and applying some editorial judgement can provide a first draft for a Boolean categorizer.

Alternatively, if a collection of already categorized content items exists, this can be analyzed to generate a first draft for a Boolean categorizer. However, pre-categorized content needs to be carefully assessed to determine if it is relevant and consistently categorized. In the case of RWJF, there was a collection of pre-categorized grant precis, but the quality and completeness of that categorization was not adequate. Among the anomalies discovered, were formulaic precis and indexing for certain Program Areas especially related to leadership development. The lesson learned is that in some cases, it may be better to build a new set of category examples, than to rely on pre-existing indexing.

1.3. Refining recall and then precision

Recall and precision tend to resolve in direct proportion to each other, meaning that generally given an increase in precision there is a comparable decrease in recall, and visa versa. The baseline from which refinements are made is very important. In the 2017 pilot project, the results had 89% precision but only 67% recall, meaning that only 11% of the results were false positives, but 33% of the total collection was not categorized at all. Looking at the trial results for each RWJF Topic shown in FIG. 5 showed that the most precise results were for Health Care Quality and Health Care Coverage, and the least precise results were for Childhood Obesity and Disease Prevention and Health Promotion. But overall, the results were impressive given that the Topics are broad and potentially ambiguous.

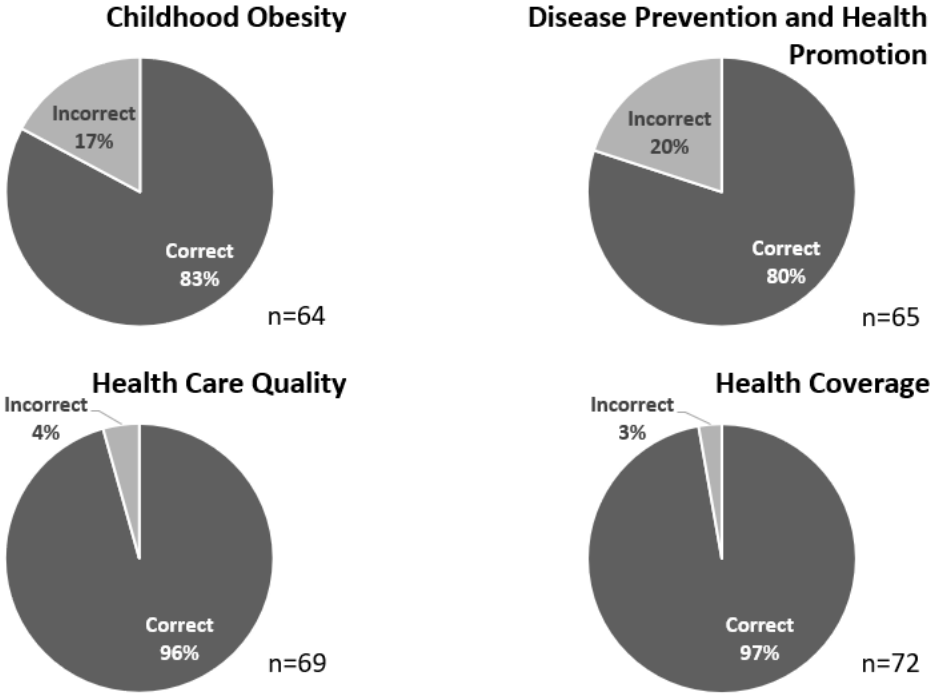


FIG. 5. 2017 pilot project results for each Broad topic.

In 2018, the process of refinement started with optimizing recall as much as possible in a first iteration of Boolean query building, and then optimizing for precision in a second iteration. While the focus of refinement is usually on precision, it is our opinion that optimizing recall is both easier and a better foundation for further refinement. This approach seeks to broaden the scope of the query and eliminate false negatives first to optimize recall, and then in a second iteration focus on the eliminating false positives to optimize precision.

1.4. Integrating text analytics into staff workflows

Beyond the development of the Boolean categorizers, developing requirements for integrating automated categorization into RWJF staff workflows raises questions about how these methods will change what people do. From the start, it was a goal to engage the Foundation's program staff directly in the process of categorizing content rather than to provide a fully automated solution to categorizing content. But this has led to some interesting discussions about who should be engaged in categorization including quality assurance. FIG. 6 shows one of the proposed workflow options for categorizing new grants.

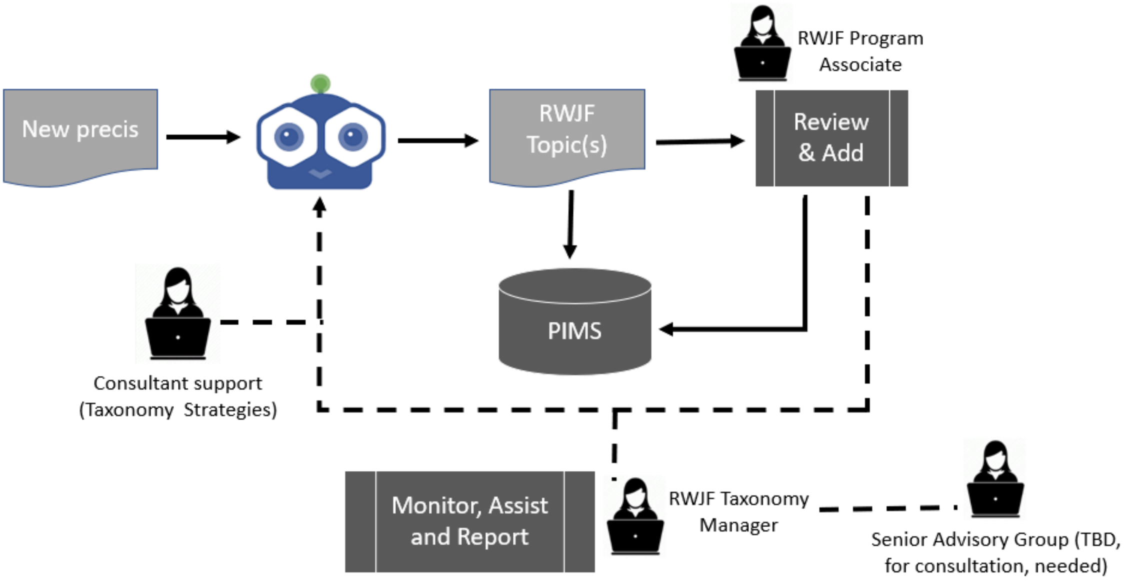

RWJF Program Associate
New precis
RWJF Topic(s)
Review & Add
PIMS
Consultant support (Taxonomy Strategies)
Monitor, Assist and Report
RWJF Taxonomy Manager
Senior Advisory Group (TBD, for consultation, needed)

FIG. 6. One proposed workflow option for categorizing new grants.

Retrospective re-categorization is planned to be a more automated process with a workflow to help users report errors, and a workflow to fix those errors and to inform users when the errors they reported have been fixed.

3. Conclusions

Working with RWJF over several years, some helpful lessons have been learned about automated categorization. These are that 1) breaking down broad topics into simple constituent queries facilitates the process of refining recall and precision by making the queries more easily understood and editable; 2) representative test collections are essential for building Boolean categorizers but even when pre-categorized collections exist they should be carefully evaluated for quality and usefulness; 3) it is effective to refine Boolean categorizers by optimizing recall before precision; and 4) automated methods should not replace staff but be a means to engage subject matter experts with content and categorization.

References

RWJF. (2018). Frequently Asked Questions. Retrieved August 13, 2018. https://www.rwjf.org/en/how-we-work/grants-explorer/faqs.html.

Lexalytics. (2018). Semantria. Retrieved August 13, 2018. https://www.lexalytics.com/semantria.

Busch, Joseph. (2018). GoToWebinar - The Current State of Automated Content Tagging: Dangers and Opportunities. July 19, 2018. Retrieved August 13, 2018. Slides - http://www.taxonomystrategies.com/wp-content/uploads/2018/01/Current%20State%20of%20Automated%20Content%20Tagging-Webinar-20180719.pdf. Script - http://www.taxonomystrategies.com/wp-content/uploads/2018/01/Current%20State%20of%20Automated%20Content%20Tagging-20180719.pdf.

Categorization Ethics: Questions about Truth, Privacy and Big Data

Presentation

Joseph A Busch
Taxonomy Strategies, USA
jbusch@taxonomystrategies.com

Keywords: bias; automated categorization; auto-classification; privacy; GDPR

Abstract

This abstract discusses issues related to the inherent bias of automated categorization caused by content collections used to build machine learning models and the impact of the General Data Protection Regulation (GDPR).

1. Introduction

Categorization is a common human behavior and it has many social implications. While categorization helps make sense of the world around us, it also affects how we perceive the world, what we like and dislike, who we feel comfortable with and who we fear. Categorization is affected by our family, culture and education. This can easily lead to classification bias where we create categories and apply them in ways that reflect bias rather than trust. (Mai) Statistical bias is caused by sampling or measurement errors. This plays out in many different contexts such as epidemiology (selection bias), the media (source omission), and machine learning (unsupervised analysis).

2. Inherent bias of automated categorization

In the October 19, 2016 ProPublica video “How Machines Learn to Be Racist,” part of a series on machine bias, Julia Angwin mentions a study where researchers analyzed 3 million words from Google news stories. The closest word associated with the phrase “black male” was “assaulted.” While the closest phrase associated with “white male” was “entitled to.” This is an illustration of the problem with an “unsupervised” analysis to identify closely associated words and phrases. It is very common to use news feeds such as Google news stories or Wikipedia as the content collection to “train” automated categorization algorithms.

How does automated categorization work? All automated categorization is based on analyzing a collection of content to identify patterns. Those patterns are transformed into examples that become “templates” for categories. There are many different scenarios that can be used to identify examples. For images, imagine a collection of examples of “cats” and “chairs.” Given enough examples, a pattern emerges that can usually determine whether an image is of a cat or a chair or not of a cat or not of a chair. FIG. 1 illustrates these image recognition rules as Boolean queries.

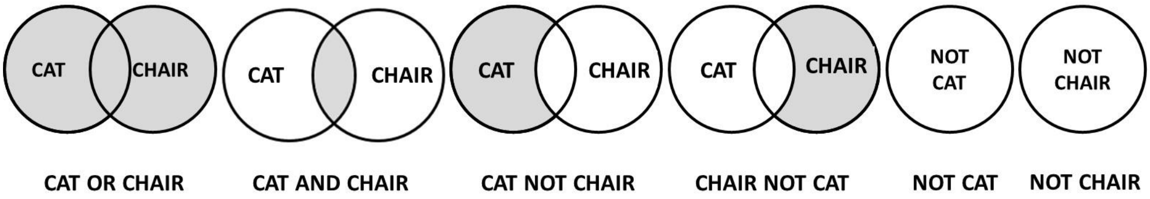
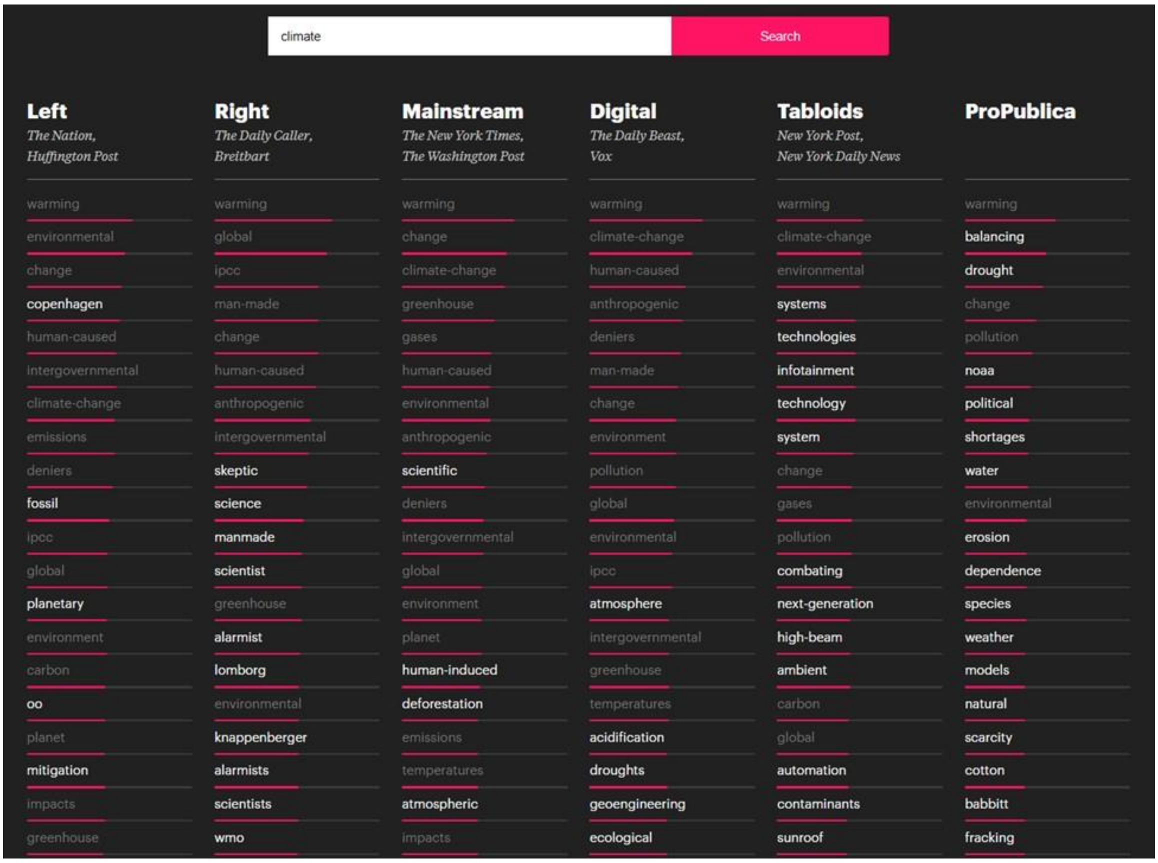


FIG. 1. Image recognition Boolean rules illustrated using Venn diagrams.

It's more complex when the collection is composed of text. In the simplest case, the text is processed using so-called natural language processing or NLP to identify nouns and noun phrases. The nouns and noun phrase occurrences and co-occurrences are counted, and then those counts are weighted based on the length of the analyzed content. Those terms with the highest weighted frequency are then used to characterize the content item. Across the content collection, other content items with similarly weighted high frequency terms are grouped together. New content items are evaluated for similarity to existing ones. Information retrieval services use these automatically generated categorizations to create feeds and make recommendations.

In the story on "How Machines Learn to Be Racist," ProPublica utilized a Google algorithm to identify synonyms (meaning closely associated nouns and noun phrases) by analyzing articles from different categories of news outlets – left, right, mainstream, digital, tabloids, and investigative. This demonstration illustrated in FIG. 2 shows how the point of view of the content collection that is processed affects the resulting list of synonyms which become the rules that define the category.


climate

Left	Right	Mainstream	Digital	Tabloids	ProPublica
The Nation, Huffington Post	*The Daily Caller, Breitbart*	*The New York Times, The Washington Post*	*The Daily Beast, Vox*	*New York Post, New York Daily News*	
warming	warming	warming	warming	warming	warming
environmental	global	change	climate-change	climate-change	balancing
change	ipcc	climate-change	human-caused	environmental	drought
copenhagen	man-made	greenhouse	anthropogenic	systems	change
human-caused	change	gases	deniers	technologies	pollution
intergovernmental	human-caused	human-caused	man-made	infotainment	noaa
climate-change	anthropogenic	environmental	change	technology	political
emissions	intergovernmental	anthropogenic	environment	system	shortages
deniers	skeptic	scientific	pollution	change	water
fossil	science	deniers	global	gases	environmental
ipcc	manmade	intergovernmental	environmental	pollution	erosion
global	scientist	global	ipcc	combating	dependence
planetary	greenhouse	environment	atmosphere	next-generation	species
environment	alarmist	planet	intergovernmental	high-beam	weather
carbon	lomborg	human-induced	greenhouse	ambient	models
oo	environmental	deforestation	temperatures	carbon	natural
planet	knappenberger	emissions	acidification	global	scarcity
mitigation	alarmists	temperatures	droughts	automation	cotton
impacts	scientists	atmospheric	geoengineering	contaminants	babbitt
greenhouse	wmo	impacts	ecological	sunroof	fracking

FIG. 2. ProPublic synonym picker illustrates how the point of view of the content affects results.

It needs to be assumed that there is an inherent bias in any collection of content that reflects discourse in a culture at a particular time, or steps need to be taken to obtain representative content—but representative of what? Bias results from models being trained on data that is historically biased. Rebecca Njeri in a 2017 blog post claims that "it is possible to intervene and address the historical biases contained in the data such that the model remains aware of gender, age and race ***without*** discriminating against or penalizing any protected classes" – (author's emphasis).

2. Impact of GDPR on automated categorization

The General Data Protection Regulation (GDPR) provides rules for protecting personally identifying information (PII), for example, the so-called "right to be forgotten." GDPR applies to processing of personal data, but not to processing of content collections in the public or published domain such as news stories or Wikipedia articles. GDPR restricts the nature of collections used for machine learning excluding anything that includes PII such as social media, customer service records, medical records, etc. Restrictions and work-arounds are already used to aggregate information in a way that obscures the PII. GDPR permits PII to be collected for specified, explicit and legitimate purposes, but does not permit further processing beyond those purposes except "for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes." (Art. 5 GDPR) Thus GDPR provides important restrictions on commercial uses of PII, even aggregated personal information, that has not been explicitly collected for a particular and personally approved purpose.

2.1. Does GDPR have an impact on classification bias?

GDPR requires that personal identifying information be accurate, and that if requested by an individual, that PII be corrected or deleted. GDPR could have an unintended impact on selection bias by allowing deletion of PII leading to incomplete or inadequate representation of a selection class.

3. Conclusions

Individuals can take responsibility for their own perceptions, misperceptions can be pointed out and sometimes changed. But categorization is often imposed on individuals from outside. For information aggregators and information analyzers, the guidelines for appropriate behavior are not always clear, nor is the responsibility for outcomes as a result of errors, bias and worse. GDPR provides some guidelines for aggregation of personal identifying information, but not on categorization bias itself. When errors and bias are commonly held, this can be reflected in the information ecology. The tipping point need not be a majority, truth or based on ethics. It's easy enough to identify cases of mis-categorization, but when should something be done about it?

References

"Art. 5 GDPR Principles relating to processing of personal data." Retrieved on August 20, 2018. https://gdpr-info.eu/art-5-gdpr/.

Dixon, Lucas and others. "Measuring and Mitigating Unintended Bias in Text Classification." Presented at: AAAI/ACM Conference on AI, Ethics, and Society (2018) Retrieved August 19, 2018. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/ab50a4205513d19233233dbdbb4d1035d7c8c6c2.pdf.

Larson, Jeff, Julia Angwin and Terry Parris Jr. "How Machines Learn to Be Racist." (October 19, 2016) Retrieved August 18, 2018. https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?

Mai, Jens-Erik. "Classification in a social world: bias and trust." 66 *Journal of Documentation* 5: 627-642 (2010) Retrieved August 19, 2018. http://jenserikmai.info/Papers/2010_Classificationinasocialworld.pdf.

Njeri, Rebecca. "How Do Machine Learning Algorithms Learn Bias?" Towards Data Science [blog] (Aug 20, 2017) Retrieved August 19, 2018. https://towardsdatascience.com/how-do-machine-learning-algorithms-learn-bias-555809a1decb.

Wikipedia. "Selection bias." Retrieved August 20, 2018. https://en.wikipedia.org/wiki/Selection_bias.

SESSION 5
Validation

Metadata quality: Generating SHACL rules from UML class diagram

Emidio Stani

Validation of a metadata application profile domain model

Mariana Curado Malta, Helena Bermúdez-Sabel, Ana Alice Baptista & Elena González-Blanco

Metadata quality: Generating SHACL rules from UML class diagrams *Presentation*

Emidio Stani
PwC, Belgium
emidio.stani@pwc.com

Keywords: model; metadata; RDF; validation; SHACL; quality; rules; Eclipse; UML

Abstract

Metadata plays a fundamental role beyond classified data, as data needs to be transformed, integrated, and transmitted. Like data, metadata needs to be harvested, standardized and validated. Metadata management processes require resources. The challenge for organizations is to make the processes more efficient, while maintaining and even increasing confidence in their data.

While RDF harvesting has already become an important step, implemented at large scale[1], there is now a need to introduce a RDF validation mechanism. However such a mechanism will depend upon the definition of RDF standards. When a standard is set, the provision of a validation service is necessary to determine if metadata complies, as for example with the HTML validation service. For example, DCAT-AP is used to describe public sector datasets in Europe; an online DCAT-AP validator[2] provides a way to validate DCAT-AP datasets.

When an organization wants to provide an RDF validation service, there are key considerations to take into account, notably the possibility for the user:

- to provide metadata to be validated in any RDF serialization, as metadata can be generated from different sources;
- to obtain the list of violations according to their severity/quality scores, allowing the user to address the most important in priority when fixing the metadata validated;
- to receive a message describing the violated rule as users might not be familiar with SPARQL or SHACL;
- and to provide/see the validation rules.

In addition, organizations have to continuously review these rules which in turn depend on the model. Thus organizations need to synchronize the rules with the current model.

Such requirements would be easily met by generating the rules automatically in order to make this process less error prone and more efficient.

A Model Driven mechanism which generates rules out of a model, is therefore a good practice since changes can be applied directly in the model and rules can automatically be generated. This approach is already a well-used technique, especially for Object Oriented Applications for models serialized, such as XML schema.

The proposed presentation will show one method to use model driven mechanism to generate automatically violations rules. Using tools for model design and model to text functions like Papyrus and Acceleo based on Eclipse, it is possible to generate SHACL constraints. A UML class diagram with stereotypes is used to describe the original metadata. Thanks to the UML stereotypes, one can then generate automatically SHACL constraints that could be then used by a SHACL

[1] **European Data Portal**, European Union. https://www.europeandataportal.eu/ Accessed on 27/08/2018

[2] **DCAT-AP validator**, Open Data Support. http://dcat-ap.semic.eu/dcat-ap_validator.html Accessed on 27/08/2018

validator. The Flemish Government has implemented a similar method by using other tools[3] and publishing an online validator called OSLO2 validator[4].

[3] **OSLO-EA-to-RDF**, Flemish Government. https://github.com/Informatievlaanderen/OSLO-EA-to-RDF Accessed on 27/08/2018

[4] **OSLO2 validator**, Flemish Government. https://data.vlaanderen.be/shacl-validator/ Accessed on 27/08/2018

Validation of a metadata application profile domain model

Mariana Curado Malta
CEOS.PP, Polythecnic of Oporto, Portugal
Algoritmi Center, UMinho, Portugal
mariana@iscap.ipp.pt

Helena Bermúdez-Sabel
LINHD - UNED, Spain
helena.bermudez@linhd.uned.es

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

Elena González-Blanco
Coverwallet, Spain
elena@coverwallet.com

Abstract

The development of Metadata Application Profiles is done in several phases. According to the Me4MAP method, one of these phases is the creation of the domain model. This paper reports the validation process of a domain model developed under the project POSTDATA - Poetry Standardization and Linked Open Data. The development of the domain model ran with two steps of construction and two of validation. The validation steps drew on the participation of specialists in European poetry and the use of real resources. On the first validation we used tables with information about resources related properties and for which the experts had to fill certain fields like, for examples, the values. The second validation used a XML framework to control the input of values in the model. The validation process allowed us to find and fix flaws in the domain model that would otherwise have been passed to the Description Set Profile and possibly would only be found after implementing the application profile in a real case.

Keywords: metadata; metadata application profile; Me4MAP

1. Introduction

The Semantic Web is an ecosystem of linked data, published, used and reused by agents related to communities of practice. The aim of these agents is to publish semantically interoperable data with data from other partners from the same community, and to profit from the open context that the ecosystem provides. In fact, the Semantic Web gives us this possibility of enriching data beyond borders and frontiers of communities since it is possible to start in a dataset "and then move through an unending set of databases which are connected not by wires but by being about the same thing" (Hawke, Herman, Archer, & Prud'hommeaux, 2013). Semantic interoperability is potentiated when data can be readily accessible with embedded information about its meaning, and it is possible through the use of common vocabularies and data models. In order to achieve maximum interoperability of its data, the development of semantic web applications requires obedience to *de jur*e and/or *de facto* standards. This implies careful and rigorous steps on the definition and design of its data and of its relationships with other data in the Web. One of the constructs that represents a semantic web data model is a Description Set Profile (DSP), which is, in turn, a component of a Metadata Application Profile (MAP). A MAP is a "generic construct for designing metadata records" (Coyle & Baker, 2009).

This paper is framed in a project funded by the European Research Grant (ERC), POSTDATA[1], which aims to provide means to make data about European poetry available as linked open data (LOD). Thus, POSTDATA is developing a MAP for the European poetry (MAP-EP). The

[1] http://postdata.linhd.es – accessed in July 31, 2018

POSTDATA work team is using the method for the development of metadata application profiles (Me4MAP) – see Curado Malta & Baptista (2013)– for its development. Me4MAP has been tested in several settings –see Curado Malta & Baptista (2017); Curado Malta, Baptista, & Parente (2015)– and this paper presents another one: European poetry provided by different institutions of the European poetry community of practice. The paper delineates how a domain model was developed in a context where non-interoperable structured data exists in 23 disperse databases that serve their own Websites, and also show in detail how this domain model was validated. The information herein presented is relevant both to the Metadata and the Digital Humanities communities. To the Metadata community because it provides a real-world example of a validation of a linked data domain model. To the Digital Humanities community because it gives information on how it is possible to create common models out of different contexts that will allow new studies across different repositories.

This paper is divided in four sections. The following section presents 1) Me4MAP and how it was used to develop the Domain Model, 2) the application domain where the MAP is being developed. Section 3 presents how the development of the Domain Model was done, presenting briefly the phases of construction of the Domain Model in the first sub-section and with more detail the phases of validation of the Domain Model in the second sub-section. The last section presents our conclusions and briefly explores future work.

2. Contextualisation

This section presents the context of this research project. The first sub-section presents the method for the development of metadata application profiles (Me4MAP) and why it is used in the development of the MAP-EP. The second sub-section introduces the European poetry community of practice as a context of the MAP-EP.

2.1 Me4MAP: a method for the development of metadata application profiles

The use of methods in any process of information systems development is important and the development of a MAP is no exception. In fact, a method introduces rigour in the process walking the developers through a path to follow and establishing which activities should be developed, when the activities may take place, how they interconnect and finally which milestones and deliverables they produce. The authors have been working in Me4MAP since 2012 and are using the process of developing MAP-EP as one more use-case to provide input for the improvement of Me4MAP.

Me4MAP presents a set of activities, organised in stages that are called the Singapore Stages. The name of the stages comes after the seminal document presented by Nilsson, Baker, & Johnston (2008). On stage S1 the Functional Requirements are defined, on stage S2 the Domain Model and on stage S3 the Description Set; these three stages are sequential and the deliverables of a previous stage feed the next stage.

As we will explain in the next paragraphs, we did not follow exactly Me4MAP for the Domain Model definition since the setting showed other possibilities.

On S1, Me4MAP defines a set of activities in order to obtain the Functional Requirements: S1.1 Definition of the Vision, S1.2 Development of the Work-Plan, S1.3 Definition of the Application Domain, S1.4 Elicitation of the high-level requirements and S1.5 Development of the Use-Case Model. The first three activities are general to all settings, the last two depend on the available resources of the setting that allow the work team to analyse the data needs of the community. And indeed, Me4MAP states that, depending on the resources available, it is possible to use other approaches to define the Functional Requirements.

Me4MAP says the Functional Requirements identified serve as input for the definition of the Domain Model. This is in fact nothing new since Me4MAP is inspired by the early stages of data modelling used in the software development processes (Curado Malta & Baptista, 2013a) –e.g. Rationale Unified Process (Kruchten, 2004). But in our work we did not elicit functional requirements. In fact, since we had already structured data in the digital repertoires available on the Web, we decided to use the database structures of these repertoires as source to define the Domain Model[2].

2.2 Community of Practice: the European poetry

The willingness of an informal group of poetry scientists, that have been working together for some years, to publish data about poetry metrics in Linked Open Data provided the perfect opportunity to propose the development of a MAP for this specific community.

The MAP-EP is being developed in the scope of the POSTDATA project, a European Research Council Starting Grant – see Curado Malta, González-Blanco, Martinez, & Del Rio (2016) for more information about the project.

The research community of poetry works with digital repertoires of poetry. A repertoire is a catalogue that gives account of the metrical and rhythmical schemes of either a poetical tradition, a period or school, gathering a corpus of poems that are defined and classified by their main characteristics. These kind of repertoires may sometimes contain the text of the poem and information related to authors, manuscripts, editions, music, and other features, all of them related to the poems (Curado Malta et al., 2016)

These repertoires exist on the Web but are not interoperable (González-Blanco & Seláf, 2014). They have real data from research projects on poetry and this data has been structured by information modellers that have built these systems without concern with the possibility of interoperability. Since their interest laid in answering the particular research questions of their project, their goal is to just serve the specific needs of the local community. The poetry scientists want now to explore new possibilities; they want to cross or compare data from different traditions that is stored in different silos of information. Also, the possibility to link the data of those silos with other resources present in the LOD ecosystem is seen as a huge opportunity to enrich the data that already exists.

3. Developing the Domain Model

The development process of defining the Domain Model was made of two well-defined moments of construction and two well-defined moments of validation (see FIG.1). Nevertheless, there were certainly less distinct tasks of validation and construction since there were informal moments of discussion with poetry scientists during local presentations in the laboratory with visitors or in meetings with all the laboratory colleagues.

The process was iterative since we defined Version 0.1[3] (DM v0.1 in FIG.1) and validate it. Out of this first validation we issued Version 0.2[4] (DM v0.2 in FIG.1). Then, in a new period of construction, we defined Version 0.3[5] (DM v0.3 in FIG.1), finally this version was validated and we issued the first stable version of the Domain Model (DM v1.0 in FIG1 – version submitted to a scientific journal, waiting for editorial decision).

3.1 Building the Domain Model

The work team identified 23 important representatives of the community of practice. Seventeen provided the database structures of the digital repertoires. We used a reverse engineering process (Müller et al., 2000) to transform the logical data models of the databases into conceptual ones.

[2] See https://goo.gl/O0mqhI for the complete set of digital repertoires used in the whole process of the Domain Model definition – accessed in July 31, 2018

[3] Available at https://doi.org/10.5281/zenodo.832885 – accessed in July 31, 2018

[4] Available at https://doi.org/10.5281/zenodo.832906 – accessed in July 31, 2018

[5] Available at http://doi.org/10.5281/zenodo.1164193 – accessed in July 31, 2018

Curado Malta, Centenera, & González-Blanco (2017) and Bermúdez-Sabel, Curado Malta, & González-Blanco (2017) expound how the Domain Model was defined having as basis conceptual models of some databases.

Regarding the repertoires for which the delegates did not provide database structures, we decided to analyse the websites identifying their informational needs since they were openly available on the Web. By informational needs we mean the data the system needs to retrieve from the database to provide the information stated on the screen and the way it combines it. Firstly, we analysed the different pages or screens of the Website and how they were linked, and then for each screen we identified each dynamic field on the screen as data to be part of the Domain Model.

We have used the digital repertoire *MedDB – Base de Datos da Lírica Profana Galego-Portuguesa*[6] to conceptualise the framework of analysis. In fact, this database was used in the first moment of construction of the Domain Model (we had access to the database structure), but our idea was to test the results of the analysis against the structure of the database to verify whether the technique used was adequate and did not miss any important data.

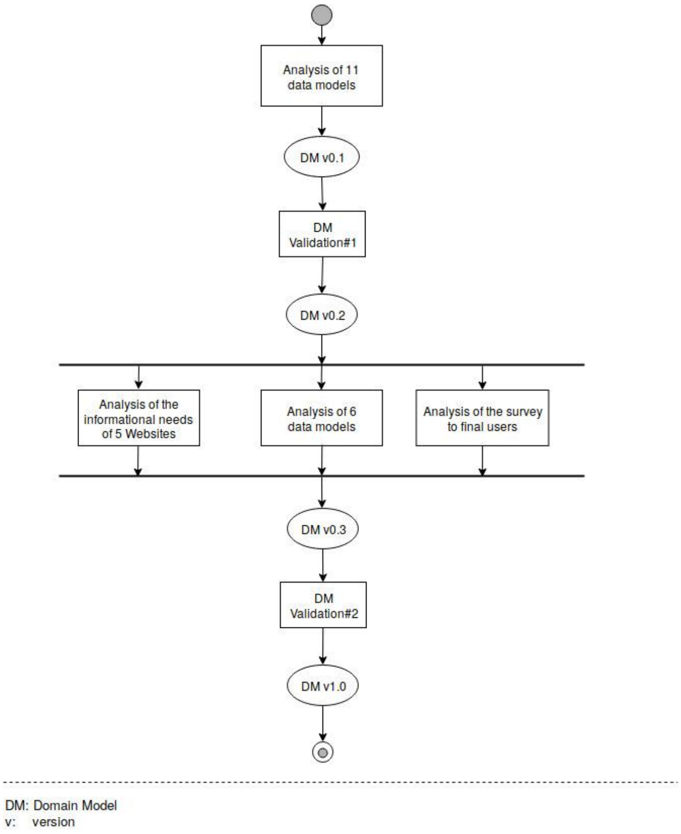

Analysis of 11 data models
DM v0.1
DM Validation#1
DM v0.2
Analysis of the informational needs of 5 Websites
Analysis of 6 data models
Analysis of the survey to final users
DM v0.3
DM Validation#2
DM v1.0
DM: Domain Model
v: version

FIG. 1. The process of development of the Domain Model of the MAP-EP

The link http://doi.org/10.5281/zenodo.1117064[7] presents the report of this analysis showing the new data needs that were introduced in Version 0.3 of the Domain Model.

The work team also made available on the Web a survey to final users of the repertoires to understand the informational needs of these users, the link

[6] http://www.cirp.gal/meddb – accessed on July 31, 2018

[7] Accessed on July 31, 2018

https://doi.org/10.5281/zenodo.1117194[8] provides the results of the survey as well as the data needs that were introduced in Version 0.3 of the Domain Model.

The next sub-section presents the activities of validation (DM Validation#1 and DM Validation#2) that were developed.

3.2. Validating the Domain Model

We implemented two moments of validation: the first one validated Version 0.1 of the Domain Model –referred as "DM Validation#1" in FIG.1, and the second moment validated Version 0.3 of the Domain Model –referred as "DM Validation#2" in FIG.1.

The paradigm behind the class diagrams is object-oriented. The paradigm behind Linked Data is property centric and one of its benefits is that "it allows anyone to extend the description of existing resources, one of the architectural principles of the Web" (Brickley & Guha, 2004). The use of modelling techniques based on two distinct paradigms may pose some problems of expressiveness and coherence between the respective models. In our case, for clarity and ease of transposition to the property centric paradigm of the Resource Description Framework (RDF), we have mapped the relations between classes as properties that have those classes as their domain and/or range. For example, a *`rel`* relationship between class *`A`* and class *`B`* would be mapped as a *`rel`* property with domain *`A`* and range *`B`*.

Domain Model Validation#1

DM Validation#1 took place in March 2017 at UNED (Madrid), the university that hosts the POSTDATA project. We invited delegates of the digital repertoires that were firstly contacted during the definition of the state-of-the-art and thus were invited to participate as stakeholders. Delegates from ten different repertoires collaborated in the discussions of the Domain Model, nine of which participated in the validation test as well since their data models were analysed during the development of the Domain Model.

Delegates were all application experts (philologists). Each delegate received as work material:

- A paper sheet with the UML class diagram of the conceptual model of its own database: this diagram included the classes of the database, the relations between the classes and the attributes of each class. It is important to note that the names of the classes were the same as the ones appearing in the Domain Model.
- A spreadsheet file with a mapping between the logical model of the database of the delegate and the conceptual model (developed in the scope of POSTDATA) of the database.
- A paper sheet with the UML class diagram of the Domain Model: The diagram included the classes and the relations between the classes. It did not include the attributes of each class for reasons of readability[9].
- A spreadsheet with i) a list of the classes of the Domain Model and description of each class, ii) a list of the attributes of the Domain Model with description. The attributes were organised by classes, and iii) a list of the relations of the classes with domain and range information.

A testing sheet was used to execute the validation[10]. This testing sheet is organised as follows:

- Each sheet (see FIG.2) has the name of a class (e.g `Opus`), on the top of the sheet there is a cell that identifies the instance of the class (value of the cell "Instance label"). The sheet can be repeated as many times as the number of instances of the class that the resource

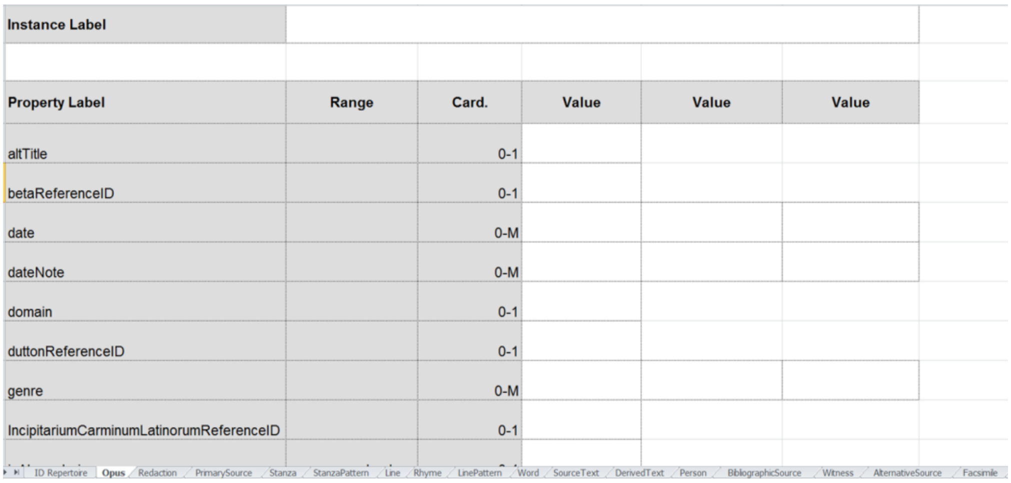
[8] Accessed on July 31, 2018

[9] Available at https://doi.org/10.5281/zenodo.437827 – accessed in July 31, 2018

[10] Available at https://doi.org/10.5281/zenodo.1226672 – accessed in 21 April, 2018

being tested has. Or, if needed and if there is space, other instances of the same class can be repeated in the same sheet.

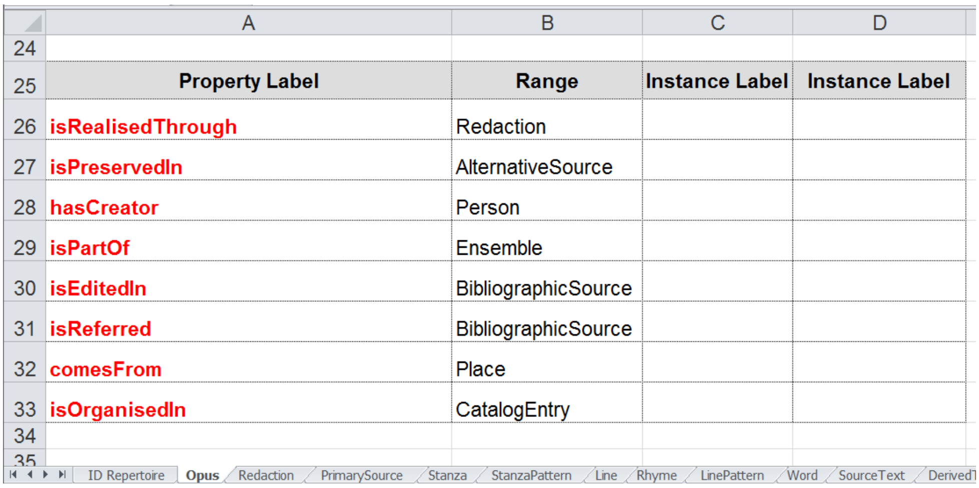
- Each sheet has a list of the attributes (column "Property Label") of the class at hand. Each line represents an attribute and has the following columns: "range" (the type of the value of the attribute, e.g. int, text, boolean), "cardinality" (how many times the attribute can be repeated) and "value" (we can have more than one column named "value", depending on the cardinality of the attribute). The cell of the columns "value" should be filled in with the information of the resource related to that attribute.
- FIG.3 presents the last part of each sheet where there is a list of the relations (Column A - "Property label") between the class at hand and other classes (Column B - "Range"). The cells in the columns "Value" should be filled with the names of the instances of the class that are the range of the relations at hand. For example, the two instances of classes `Opus1` and `Redaction1` relate the following way: `Opus1—isRealisedThrough—Redaction1`. This is made explicit by filling in:
  - sheet "Opus" (see FIG.2), the cell "Instance label" with the value `Opus1` and,
  - the same sheet "Opus" (see FIG.3), the cell C26 with the value `Redaction1`.



Instance Label

Property Label	Range	Card.	Value	Value	Value
altTitle		0-1			
betaReferenceID		0-1			
date		0-M			
dateNote		0-M			
domain		0-1			
duttonReferenceID		0-1			
genre		0-M			
IncipitariumCarminumLatinorumReferenceID		0-1			

FIG. 2. An excerpt of the test sheet: list of some attributes of the class Opus

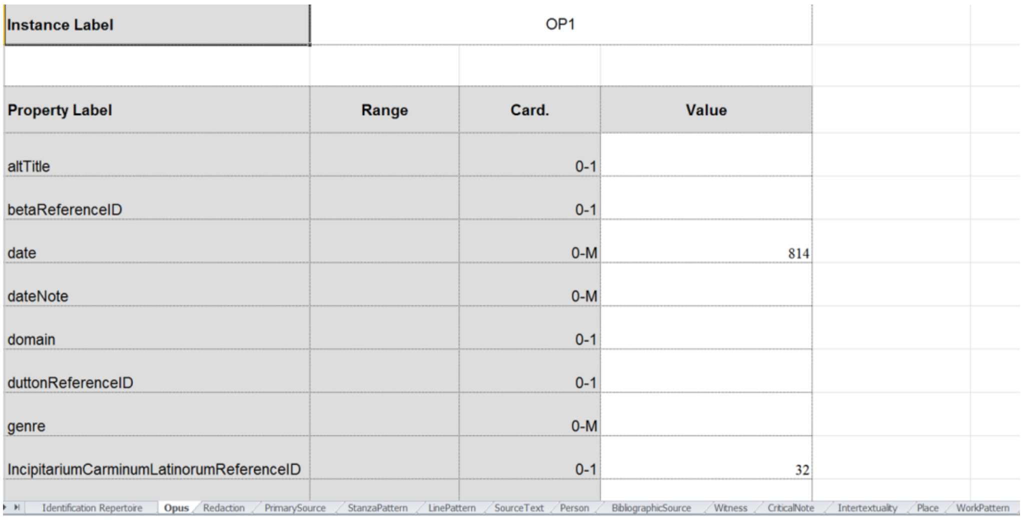


	A	B	C	D
24				
25	Property Label	Range	Instance Label	Instance Label
26	isRealisedThrough	Redaction		
27	isPreservedIn	AlternativeSource		
28	hasCreator	Person		
29	isPartOf	Ensemble		
30	isEditedIn	BibliographicSource		
31	isReferred	BibliographicSource		
32	comesFrom	Place		
33	isOrganisedIn	CatalogEntry		
34				
35				

FIG. 3. An excerpt of the test sheet: list of some relations of the class Opus

Before doing the hands-on session of validation, a testing sheet with an example of testing was given to the delegates and explained[11] for the delegates to understand the aim of the session. The example given used a resource sample from the repertoire *Corpus Rhythmorum Musicum*[12], one of the repertoires used to build the Domain Model. FIG.4 shows an excerpt of the validation example:

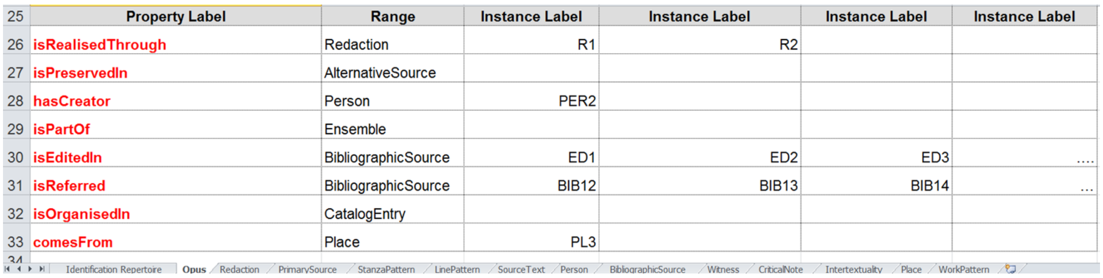
- There is an instance of the class `Opus` named OP1;
- Attributes of `OP1`, e.g. the date of creation (value: year 814) and the Reference ID of the catalogue Incipiarium Carminum Latinorum (value: 32);
- `OP1` relates to a certain number of other instances of classes (see FIG.5): `OP1 isRealisedThroug` R1, `OP1 isRealisedThrough R2`, `OP1 hasCreator PER2`, etc. All these instances of classes (`R1`, `R2`, `PER2`, etc.) have sheets where their attributes are defined;
- FIG.6 presents several instances of the class `Person` where `PER2` (the author of `OP1`) in the attribute `name` has the value "anonymous", meaning that `OP1` has an anonymous author.



Instance Label	OP1		
Property Label	**Range**	**Card.**	**Value**
altTitle		0-1	
betaReferenceID		0-1	
date		0-M	814
dateNote		0-M	
domain		0-1	
duttonReferenceID		0-1	
genre		0-M	
IncipitariumCarminumLatinorumReferenceID		0-1	32

Identification Repertoire | Opus | Redaction | PrimarySource | StanzaPattern | LinePattern | SourceText | Person | BibliographicSource | Witness | CriticalNote | Intertextuality | Place | WorkPattern

FIG. 4. An excerpt of the validation example: instance OP1 of the class Opus



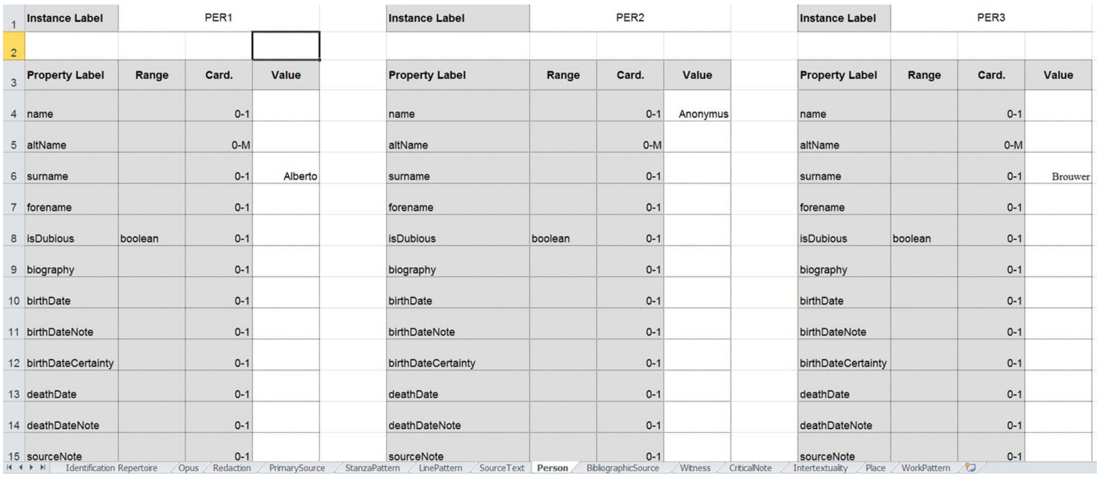
25	Property Label	Range	Instance Label	Instance Label	Instance Label	Instance Label
26	isRealisedThrough	Redaction	R1	R2		
27	isPreservedIn	AlternativeSource				
28	hasCreator	Person	PER2			
29	isPartOf	Ensemble				
30	isEditedIn	BibliographicSource	ED1	ED2	ED3	....
31	isReferred	BibliographicSource	BIB12	BIB13	BIB14	...
32	isOrganisedIn	CatalogEntry				
33	comesFrom	Place	PL3			
34						

Identification Repertoire | Opus | Redaction | PrimarySource | StanzaPattern | LinePattern | SourceText | Person | BibliographicSource | Witness | CriticalNote | Intertextuality | Place | WorkPattern

FIG. 5. An excerpt of the validation example: instance OP1 of the class Opus and the relations with other instances of classes

[11] See https://doi.org/10.5281/zenodo.1226672 –accessed in July 31, 2018 – to download the file

[12] See http://www.corimu.unisi.it/ – accessed in July 31, 2018



1	Instance Label	PER1			Instance Label	PER2			Instance Label	PER3		
2												
3	Property Label	Range	Card.	Value	Property Label	Range	Card.	Value	Property Label	Range	Card.	Value
4	name		0-1		name		0-1	Anonymus	name		0-1	
5	altName		0-M		altName		0-M		altName		0-M	
6	surname		0-1	Alberto	surname		0-1		surname		0-1	Brouwer
7	forename		0-1		forename		0-1		forename		0-1	
8	isDubious	boolean	0-1		isDubious	boolean	0-1		isDubious	boolean	0-1	
9	biography		0-1		biography		0-1		biography		0-1	
10	birthDate		0-1		birthDate		0-1		birthDate		0-1	
11	birthDateNote		0-1		birthDateNote		0-1		birthDateNote		0-1	
12	birthDateCertainty		0-1		birthDateCertainty		0-1		birthDateCertainty		0-1	
13	deathDate		0-1		deathDate		0-1		deathDate		0-1	
14	deathDateNote		0-1		deathDateNote		0-1		deathDateNote		0-1	
15	sourceNote		0-1		sourceNote		0-1		sourceNote		0-1	

FIG. 6. An excerpt of the validation example: instances of the concept “Person” of the resource being described

We asked the delegates to choose some resources from their own digital repertoires and fill in the validation sheet with the correspondent values.

During the process of validation, we asked the delegates to register the issues that arose during the validation tests in the validation sheet. Also, at the end of the workshop we asked the delegates to upload the file(s) with the validation tests to a server in order to be analysed later by the work team. The delegates were also asked to fill in a form with the following questions:

- Could you describe all your data with the available elements? If not, please refer the difficulties.
- Did you have any difficulty in particular to describe your data? Were there any ambiguities?
- Is there anything else you want to add?

The work team used all the inputs given by the delegates to issue a Version 0.2 of the Domain Model.

Domain Model Validation#2

The DM Validation#2 was done on Version 0.3 of the Domain Model. In similarity to the previous process of analysis of the informational needs of the Websites, the digital repertoire *MedDB – Base de Datos da Lírica Profana Galego-Portuguesa* was used to conceptualise the framework of validation. After that, we have identified resources from digital repertoires that were not part of the 17 repertoires used as sources during the processes of construction, this way we could address at a certain point the general scope of the Domain Model. By “general scope” we mean that we expect this Domain Model to serve other contexts, in the same community of practice, then the ones used to create it.

This validation was done mainly by a master student of philology that did not participate in the processes of construction of the Domain Model. By using an external person to the team we wanted to give total freedom of interpretation of the model to see if again the Domain Model could respond to the needs of the community. This student was helped by members of the team, nevertheless we tried not to introduce any bias on the use-cases building.

The DM Validation#2 consisted in using real resources from the GUIs of the databases and, with that data, populate the Domain Model Version 0.3. For this work we created:

1. A description of the Domain Model in XML[13];
2. Schema files for the use-cases that validate their contents against the DM[14].

Besides the repertoire used as base example for each process, the aforementioned *MedDB*, we selected five different poetry projects and randomly chose, at least, one poetic resource from each one them. In total, we built nine use-cases.

The modelling of the use-cases consisted in describing the resource in XML using the classes, attributes and relations of the Domain Model. The schema file restricted both the classes and the different attributes and relations, so any elements that were not contained in the Domain Model could not be added. In addition, it also prevented the repetition of labels that identified the different instances of each class so to avoid ambiguities. This schema also controlled the relations between the different instances of class: except for the instance of class Opus, every instance of any other class had to be the range of at least one relation.

The construction of the use-cases affected the contents of the DM. Whenever an informational need not previously considered was detected, the elements required for enabling its modelling were added to the Domain Model so we had an updated version to validate against the use-cases. This means that the XML provided as representative of Version 0.3 represents a previous stage of the Domain Model than the use-cases.

With the information retrieved from the construction of the use cases, we created a report, organised by digital repertoire. This report is available on http://doi.org/10.5281/zenodo.1164854[15].

Out of Validation#2 we issued the first stable version of the Domain Model for European Poetry that is to be published in a scientific journal (waiting for editorial decision).

4. Conclusions and Future Work

A metadata Application profile (MAP) is a construct of the semantic web that enhances interoperability (Nilsson et al., 2008). When a community of practice publishes linked open data (LOD) in the semantic web using as reference the MAP of the community, all the data from its datasets will be ready to be used and combined automatically since they have exactly the same structure. Adding to this, if the developers of the MAP followed good practices while defining it, i. e., used standard vocabularies of the semantic web and referenced resources of other datasets inside borders of the same community or even outside, these data will be much enriched. An informal group of philologists, delegates of digital repertoires of European poetry, understood that they could profit from these possibilities. The POSTDATA project, financed by a European Research Council (ERC) Grant, started two years ago with the aim (among other goals) of providing means for this informal group –and later any organisation of the same community of practice– to publish LOD about European poetry. To achieve this goal, the POSTDATA work team decided to develop a MAP for the European poetry (MAP-EP) using Me4MAP, a method for the development of MAPs. This paper presents the work developed during the definition of the Domain Model of this MAP-EP, more specifically presents how the validation of the Domain Model was done. The process followed during the building of the Domain Model for European poetry reveals the importance of validation, hence the upgrade of version that each validation moment caused.

This validation included two steps: 1) The first moment had the aim to validate Domain Model version 0.1. This occurred in a workshop with the informal group just referred where they tested

[13] See https://github.com/postdataproject/Domain-Model-v.0.3/tree/master/domain-model – accessed in July 31, 2018 – for the XML file with the description of the Domain Model and the related schemas)

[14] See https://github.com/postdataproject/Domain-Model-v.0.3/tree/master/use-cases – accessed in July 31, 2018 – for the XML files of the use-cases and related schemas

[15] Accessed on July 31, 2018

the model using real resources from their own databases. This group was guided to populate a testing file with information from the chosen resources. The file was organised in such a way that it reproduced the structure of the Domain Model in worksheets of a spreadsheet; 2) the second moment had the aim to validate Domain Model version 0.3. This validation was an activity that consisted in using a set of use-cases, resources of other digital repertoires –other than the used in the building of the model– and feeding XML files with the information from the resources. The XML files were structured in a way that reproduced the Domain Model, and we used a XML framework to validate in real-time the values introduced to avoid any errors. This last validation activity resulted in version 1.0 of the Domain Model.

The first stable version of the Domain Model for the European poetry is a milestone of the whole process of developing the MAP-EP. The POSTDATA work team is now continuing the development of MAP-EP. The current work is focusing in aligning each concept of the DM (either class, attribute or relation) with the RDF vocabulary term that best describes it, as well as developing vocabulary encoding schemes to constrain certain properties.

This activity of developing a Domain Model in the framework of a MAP development was the opportunity for Me4MAP researchers to test the method in a new setting not tested before. Me4MAP was developed following a Design Science Research methodological approach –see Hevner (2007). During its development, the method was tested using an experimental situation with a worldwide group of the Social and Solidarity Economy (SSE) to collaboratively build a MAP for the Web Based Information Systems of the SSE community (Curado Malta, 2014; Curado Malta, Baptista, & Parente, 2015). Me4MAP researchers think that Me4MAP may be adequate in a context similar to the one used in the SSE community, but it needs validation in different settings. In fact the question of generalisability needs to be addressed as well as the limits of the Me4MAP applicability. This is why this new use-case of Me4MAP application is being monitored. The work described in this paper will be subject of reflection in order to give input for the improvement of Me4MAP.

Final Note: The authors are sorted in descending order according to their contribution to research and writing.

Acknowledgements

The authors would like to thank the delegates of the analysed repertoires for their availability and willingness to share information and to discuss issues related to their projects with the POSTDATA team.

The authors would like thank Patricia Garrido Teixeira for her work on the use-cases and Clara Martínez (LINHD-UNED) for her help on the metrical analysis of the digital repertoires.

The research presented in this paper was developed with the support of the Starting Grant research project: Poetry Standardization and Linked Open Data - POSTDATA (ERC-2015-STG-679528), funded by the European Research Council (ERC) under the European Union's Horizon2020 research and innovation programme.

Part of this work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

References

Bermúdez-Sabel, H., Curado Malta, M., & González-Blanco, E. (2017). Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts. In Language, Data, and Knowledge (pp. 156–165). Springer, Cham. https://doi.org/10.1007/978-3-319-59888-8_14

Brickley, D., & Guha, R. V. (2004, February). RDF Vocabulary Description Language 1.0: RDF Schema. Retrieved 7 May 2018, from http://www.w3.org/TR/rdf-schema/

Coyle, K., & Baker, T. (2009). DCMI: Guidelines for Dublin Core Application Profiles (Working Draft). Retrieved 15 January 2018, from http://dublincore.org/documents/profile-guidelines/

Curado Malta, M. (2014, July 16). Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web Semântica. Universidade do Minho, Guimarães, Portugal. Retrieved from http://hdl.handle.net/1822/30262

Curado Malta, M., & Baptista, A. A. (2013). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In Proceedings on International Conference on Dublin Core and Metadata Applications (pp. 90–103). Lisbon, POrtugal: Dublin Core Metadata Initiative. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3674

Curado Malta, M., & Baptista, A. A. (2017). The Development process of a Metadata Application Profile for the Social and Solidarity Economy: Computer Science & IT Book Chapter | IGI Global. In Developing Metadata Application Profiles (pp. 98–117). IGI Global. Retrieved from https://www.igi-global.com/chapter/the-development-process-of-a-metadata-application-profile-for-the-social-and-solidarity-economy/175868

Curado Malta, M., Baptista, A. A., & Parente, C. (2015). A DCAP for the Social and Solidarity Economy. In Proceeedings of the International Conference on Dublin Core and Metadata Applications (pp. 20–29). Lisbon, Portugal: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2015/paper/view/372

Curado Malta, M., Centenera, P., & González-Blanco, E. (2017). Using Reverse Engineering to Define a Domain Model: The Case of the Development of a Metadata Application Profile for European Poetry. In Developing Metadata Application Profiles (pp. 146–180). Retrieved from https://www.igi-global.com/chapter/using-reverse-engineering-to-define-a-domain-model/175870

Curado Malta, M., González-Blanco, E., Martínez, C., & Del Rio, G. (2016). Digital repertoires of poetry metrics: towards a Linked Open Data ecosystem. In Proceedings of the First Workshop on Digital Humanities and Digital Curation co-located with the 10th Conference on Metadata and Semantics Research (Vol. 1764, pp. 1–11). Gröningen, Germany: CEUR Workshop Proceedings.

González-Blanco, E., & Seláf, L. (2014). Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires. Humanitats a La Xarxa: Món Medieval/Humanities on the Web: The Medieval World, 321–332.

Hawke, S., Herman, I., Archer, P., & Prud'hommeaux, E. (2013). W3C Semantic web activity. Retrieved 4 May 2018, from https://www.w3.org/2001/sw/

Hevner, A. R. (2007). The three cycle view of design science research. Scandinavian Journal of Information Systems, 19(2), 87–92.

Kruchten, P. (2004). The rational unified process: an introduction (3rd ed.). Boston, MA, USA: Addison-Wesley Professional.

Müller, H. A., Jahnke, J. H., Smith, D. B., Storey, M.-A., Tilley, S. R., & Wong, K. (2000). Reverse Engineering: A Roadmap. In Proceedings of the Conference on The Future of Software Engineering (pp. 47–60). New York, NY, USA: ACM. https://doi.org/10.1145/336512.336526

Nilsson, M., Baker, T., & Johnston, P. (2008). DCMI: Singapore Framework for Dublin Core Application Profiles. Retrieved 20 April 2018, from http://dublincore.org/documents/profile-guidelines/

SESSION 6
Application Profiles

Modeling and application profiles in the Art and Rare Materials BIBFRAME Ontology Extension

Jason Kovari
Cornell University, USA
jak473@cornell.edu

Melanie Wacker
Columbia University, USA
mw2064@columbia.edu

Huda Khan
Cornell University, USA
hjk54@cornell.edu

Steven Folsom
Cornell University, USA
sf433@cornell.edu

Keywords: linked data; BIBFRAME; application profiles; art objects; rare materials; ontologies; semantic applications

Abstract

Between April 2016 and July 2018, the Art Libraries Society of North America's Cataloging Advisory Committee (CAC) and the RBMS Bibliographic Standards Committee (BSC) collaborated with the Andrew W. Mellon Foundation funded Linked Data for Production (LD4P) project on the Art and Rare Materials BIBFRAME Ontology Extension (ARM). The motivation for this effort stems from BIBFRAME purposefully under-defining modeling for realms considered outside of core bibliographic description, expecting specialized communities to build extension ontologies.

In this context, ARM facilitates the descriptive needs of the art and rare materials communities; modeling includes areas such as exhibitions, materials, measurements, physical condition and other realms, as well. For each area, narrative recommendations documents were written that included use cases, diagrams and terms from relevant ontologies. Further, OWL ontologies files were developed for both the newly-defined ARM terms as well as target ontologies expected to be used alongside ARM, as defined in the aligned recommendation documents and SHACL application profiles. ARM ontology files were divided into four modularized ontologies: Core, which includes all ARM terms not identified for other ontology files; Award, which includes all terms relevant to the description of awards received by an agent or other resource; Custodial History, which includes terms relevant to the provenance or custodial history of an object; and Measurement, which includes terms relevant to the description of measurements of an object. The modularized approach was selected to encourage reuse of models by communities other than art and rare collections as well as communities not using BIBFRAME as their core modeling. These ontologies were published to https://w3id.org/, a lightweight solution affording publishing these ontology files without developing infrastructure while communities of practice consider long-term maintenance, hosting and governance.

In February 2018, development effort shifted focus to a Shapes Constraint Language (SHACL) application profiles for art resources as well as a SHACL application profile for rare monographs. SHACL is an RDF-based W3C recommendation; as such, it can be represented as linked data and easily made available for reuse and extension by other communities. SHACL affords both validation and non-validation property shapes. The non-validating property shape characteristics available in SHACL benefited the ARM project in that the primary goal in developing application profiles was to create forms within an editing environment.

These application profiles were used to define forms and display for the cataloging environment in VitroLib, an RDF-based, ontology agnostic cataloging tool developed as part of the Linked Data for Libraries - Labs project. VitroLib customization requires idiosyncratic development of property groups and custom forms. As such, the ARM SHACL files were translated into code understood by VitroLib; Ideally,

future editor environments will use specifications like SHACL natively. Implementing these applications profiles in VitroLib afforded catalogers the ability to test the ARM modeling in a real-world environment, providing feedback to the project for potential future development through two workshops held June 2018.

LD4P support for ARM concluded July 2018. As of September 2018, the standards bodies of multiple archival, art, rare and special collections library professional organizations are actively discussing how best to continue development of ARM; the authors of this paper believe that this will be determined shortly following DCMI 2018.

Acknowledgements

The authors would like to acknowledge the Andrew W. Mellon Foundation for their generous support as well as all colleagues that contributed to ARM development.

References

Khan, H., Rayle, E.L. and Younes, R (2017). VitroLib: From an ontology and instance editor to a linked data cataloging editor. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2017. Retrieved August 25, 2018, from http://dcevents.dublincore.org/IntConf/dc-2017/paper/view/507

Linked Data for Libraries - Labs (2018). The VitroLib Metadata Editor. Retrieved August 25, 2018 from https://wiki.duraspace.org/x/0rmdB

Linked Data for Production (2018). Art and Rare Materials (ARM) BIBFRAME Ontology Extension GitHub repository. Retrieved August 25, 2018, from https://github.com/LD4P/arm

Linked Data for Production (2018). ARM Awards Ontology. Retrieved August 25, 2018, from https://w3id.org/arm/award/ontology/0.1/award.html

Linked Data for Production (2018). ARM Core Ontology. Retrieved August 25, 2018, from https://ld4p.github.io/arm/core/ontology/0.1/core.html

Linked Data for Production (2018). ARM Custodial History Ontology. Retrieved August 25, 2018, from https://w3id.org/arm/custodial_history/ontology/0.1/custodial_history.html

Linked Data for Production (2018). ARM Measurements Ontology. Retrieved August 25, 2018, from https://w3id.org/arm/measurement/ontology/0.1/measurement.html

Linked Data for Production (2018). ARM Recommendation Documents. Retrieved August 25, 2018, from https://github.com/LD4P/arm/tree/master/modeling_recommendations

Linked Data for Production (2018). ARM SHACL application profile : art objects. Retrieved August 25, 2018, from https://github.com/LD4P/arm/tree/master/application_profiles/art/shacl

Linked Data for Production (2018). ARM SHACL application profile : rare monographs. Retrieved August 25, 2018, from https://github.com/LD4P/arm/tree/master/application_profiles/raremat_monograph/shacl

W3C (2017). Shapes Constraint Language. Retrieved August 25, 2018, from https://www.w3.org/TR/2017/REC-shacl-20170720/

Developing a Metadata Application Profile for the Daily Hire Labour Presentation

Sangeeta Sen
Dept. of Computer Sc. and Engg.
NIT Durgapur
Durgapur, India
sangeetaaec@gmail.com

Nishat Raza
Dept. of Computer Sc. and Engg.
NIT Durgapur
Durgapur, India
nishat.nitdgp.it.2018@gmail.com

Animesh Dutta
Dept. of Computer Sc. and Engg.
NIT Durgapur
Durgapur, India
animeshnit@gmail.com

Mariana Curado Malta
CEOS.PP, Polythecnic of Oporto, Portugal
ALGORITMI Center, UMinho, Portugal
mariana@iscap.ipp.pt

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

Keywords: application profile development, Me4MAP, use cases, functional requirements, India informal sector

Abstract

EMPOWER SSE is a Foundation for Science and Technology (Fundação para a Ciência e Tecnologia - FCT, Portugal) and Department of Science & Technology (DST, India) financed research project that aims to use a Linked Open Data approach to empower Social and Solidarity Economy (SSE) Agents. It is a collaborative project between India and Portugal that is focused on defining a Linked Open Data framework to consolidate players of the informal sector, enabling a paradigm shift. The Indian economy can be mainly categorized into two sectors: formal and informal (Rada, C. (2009)). The informal sector differs from the formal as it is an unorganized sector and comprised of economic activities that are not covered by formal arrangements such as taxation, labour protections, minimum wage regulations, unemployment benefits, or documentation. The informal sector is mainly made of skilled people that follow their family job traditions, sometimes they are not even formally trained. The major economy in India depends on the skilled labour of this informal sector such e.g. farmers, electricians, food production, and small-scale industries (Kalyani, (2016)). This sector struggles with the lack of information, data sharing needs and interoperability issues across systems and organisational boundaries. In fact, this sector does not have any visibility to the society and therefore does not have the possibility to do business, as most of the agents of this sector do not reach the end of the chain. This blocks them from getting proper exposure and a better livelihood. Here agents can be job seekers or job providers.

The possibility to publish Linked Open Data (LOD) to portray the skills of these workers of the informal sector will help them to be more visible in the digital world, opening the possibility of other technological agents to build software systems that are fed by this data. In fact, the LOD paradigm will provide a way to establish the connection between skilled labour and common people. This possibility will also allow the informal sector to contribute to the development of India.

The Semantic Web is an ecosystem that enhances interoperability, enabling though scalability by opening the possibility to cross information between LOD resources on the Semantic Web cloud. The possibilities of inference over this LOD will eventually open new knowledge to raise new awareness on policy makers.

The Linked Open Data eco-system allows to connect resources from all datasets available as LOD. In order to be semantically interoperable with people with the same or similar skills and people requiring services, the datasets need to be structured following common models and using standard RDF vocabularies. A metadata application profile (MAP) is a "generic construct for designing metadata records" (Coyle & Baker, 2009), it is a model used to identify the metadata elements and the constraints over the data for a particular domain or application. According to Nilsson, Baker, & Johnston (2008), a MAP is a construct that enhances interoperability.

EMPOWER SSE is developing a MAP (DH-MAP: Daily Hire-Metadata Application Profile) for the informal sector considering that this sector needs to be interoperable with the SSE community (see Curado Malta, Baptista, & Parente, (2015)) and the formal sector. The goal is to develop a MAP that describes the workers (or groups of workers) and their interactions with the job provider. We also intend to develop a software application to handle this data. The main goal of this application is to provide a way to place an appointment between the employer (job provider) and a skilled worker (job seeker) in the informal sector. Job providers first place their order, or need of a certain work, in the system. According to the answer of the system, they book a worker or worker group and make an appointment. After job completion, job providers rate the worker. All the data will be available as LOD in a triplestore for use and reuse by people and machines.

The development of the MAP-DH follows Me4MAP (see Curado Malta & Baptista (2013)). Me4MAP defines as first step the elicitation of the functional requirements. We have built a use-case model to identify the functional requirements. The second step defined by Me4MAP is to build a Domain Model. This model is defined with the information that comes from the functional requirements. The third step is the presentation of the Description Set Profile (DSP). To achieve such a goal, Me4MAP states that there is the need to build a constraints matrix which is the matching of an RDF vocabulary term with each property of the domain model. It also provides information about the constraints (cardinality, syntax encoding schemes, vocabulary encoding schemes, domain, range). We are currently working in the constraints matrix.

The goal of the presentation is to discuss the the draft version of the milestones of the MAP done until now (such as use cases, functional requirements, domain model and constraints matrix) and to receive feedback from the metadata community.

Acknowledgements

This work has been developed under the "EMPOWER SSE: A Linked Open Data Framework for the Empowerment of Social and Solidarity Economy Agents" project, in the scope of the Indoportugal Intergovernmental Programme of Cooperation in Science and Technology. Department of Science & Technology (DST, India). Ref No. DST/INTPORTUGAL/P- 06/2017

This work has been partially supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

References

Coyle, K., & Baker, T. (2009). DCMI: Guidelines for Dublin Core Application Profiles (Working Draft). Retrieved 15 January 2018, from http://dublincore.org/documents/profile-guidelines/

Curado Malta, M., & Baptista, A. A. (2013). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In Proceedings on International Conference on Dublin Core and Metadata Applications (pp. 90–103). Lisbon, Portugal: Dublin Core Metadata Initiative. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3674

Curado Malta, M., Baptista, A. A., & Parente, C. (2015). A DCAP for the Social and Solidarity Economy. In Proceeedings of the International Conference on Dublin Core and Metadata Applications (pp. 20–29). Lisbon, Portugal: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2015/paper/view/372

Kalyani, M. (2016). Indian informal sector: an analysis. International Journal of Managerial Studies and Research (IJMSR), vol 4. 2016.

Nilsson, M., Baker, T., & Johnston, P. (2008). DCMI: Singapore Framework for Dublin Core Application Profiles. Retrieved 20 April 2018, from http://dublincore.org/documents/profile-guidelines/

Rada, C. (2009). Formal and informal sectors in China and India: An accounting-based approach (No. 2009-02). Working Paper, University of Utah, Department of Economics.

SESSION 7
Models

Research data management in the field of Ecology: an overview

Cristiana Alves, João Aguiar Castro, João Pradinho Honrado & Ângela Lomba

Metadata Models for Organizing Digital Archives on the Web: Metadata-Centric Projects at Tsukuba and Lessons Learned

Shigeo Sugimoto, Senan Kiryakos, Chiranthi Wijesundara, Winda Monika, Tetsuya Mihara & Mitsuharu Nagamori

Research data management in the field of Ecology: an overview

Cristiana Alves
CIBIO/InBIO, Portugal
cristianamaiaalves@gmail.com

João Aguiar Castro
INESC TEC, Portugal
joaoaguiarcastro@gmail.com

Cristina Ribeiro
INESC TEC, FEUP, Portugal
mcr@fe.up.pt

João Pradinho Honrado
CIBIO/InBIO, FCUP Portugal
jhonrado@fc.up.pt

Angela Lomba
CIBIO/InBIO, Portugal
angelalomba@fc.up.pt

Abstract

The diversity of research topics and resulting datasets in the field of Ecology (the scientific study of ecological systems and their biodiversity) has grown in parallel with developments in research data management. Based on a meta-analysis performed on 93 scientific references, this paper presents a comprehensive overview of the use of metadata tools in the Ecology domain through time. Overall, 40 metadata tools were found to be either referred or used by the research community from 1997 to 2018. In the same period, 50 different initiatives in ecology and biodiversity research were conceptualized and implemented to promote effective data sharing in the community. A relevant concern that stems from this analysis is the need to establish simple methods to promote data interoperability and reuse, so far limited by the production of metadata according to different standards. With this study, we also highlight challenges and perspectives in research data management in the domain of Ecology towards best practice guidelines.

Keywords: Biodiversity, Ecology, Research Data Management, Metadata tools, Literature review

1. Introduction

Ecology (the scientific study of ecological systems and of the biodiversity therein) is a challenging research community from the perspective of data management. Ecological and biodiversity data have been collected by researchers individually or as part of research teams, in the context of specific research questions and projects. Underlying data collection through time have been research topics such as the dynamics of specific habitats; the distribution and abundance of species; patterns and changes of environmental conditions; the processes that influence biological populations, communities, and ecosystems; and anthropogenic drivers of these processes (Berkley et al., 2009). Ecological and biodiversity data are collected by researchers using a wide variety of protocols tailored to address very diverse topics ranging from marine/terrestrial ecosystems to species distribution or genetics (Berkley, Jones, Bojilova, & Higgins, 2001). As a result, heterogeneous data are stored as independent datasets or databases that are dispersed throughout the research data facilities managed by ecological research communities. At the same time, to answer the multiple research questions, the need to share, describe and deposit data is a concern for many biodiversity researchers around the world.

Researchers are increasingly expected to take several measures regarding research data management (RDM), namely to comply with mandates that promote actions regarding data organization, sharing and publication. Benefits such as obtaining credit via citation or improving research workflows through collaboration may also encourage researchers to disseminate their data. Yet, availability of research data is not the same as existence of fit-for-reuse data (Tani, Candela, & Castelli, 2013). It depends, among other aspects, on specific metadata being provided to researchers so they can understand the data being accessed and evaluate their suitability. The inability to provide auxiliary information to contextualize research data is a practical impediment on data reuse (Thanos, 2017). In order to promote quality metadata, the European Commission (EC) is defining the principles to make data Findable, Accessible, Interoperable and Reusable, through the Guidelines on FAIR Data Management in Horizon 2020 (European Commission. Guidelines on FAIR Data Management in Horizon 2020., 2016).

The Research Data Alliance Metadata Standards Directory Working Group set out a metadata standards directory (Ball, Greenberg, Jeffery, & Koskela, 2016) for specific domains (Life Sciences, Engineering, Social and Behavioural Sciences) and for more general purposes. Nevertheless, the lack of resources in the long-tail of science (Heidorn, 2008) prompts researchers themselves to become active RDM stakeholders during the lifetime of projects, mostly to comply with funder or institutional policies and meet standards for good practice (Lyon, 2007). This means that research projects that do not have dedicated human resources to create standard-compliant metadata records place additional effort on researchers in the description of their data. Moreover, most standards are developed to describe data only at the end of the research workflow, with complex requirements that prevent researchers from adopting them consistently (Qin & Li, 2013). An evaluation of several metadata standards show that, although the flexibility to add new elements or modules to address community needs is a common objective in the development of scientific metadata standards, simplicity and sufficiency are not a top priority among them (Willis, Greenberg and White, 2012). Nevertheless, these features are likely to encourage researchers to describe their data, by making the process as easy as possible and focus on a minimal set of relevant metadata elements for the researchers to fill in.

Researchers are already metadata producers, yet in an ad-hoc sense and to fulfil specific, immediate needs (Mayernik, 2011). If provided with adequate tools, they are also more apt to describe context than information professionals. A promising path is to adopt metadata solutions that are tailor-made for researchers and their projects and can promote data reuse. Application Profiles, following the Singapore Framework logic of combining different standards components (Nilsson, 2008), are a practical implementation scenario to meet community-oriented metadata needs, offering the desirable flexibility but also enabling simplicity and sufficiency. The Minimum Information Framework, proposed in the geobiology community, for systematic documentation of sampling processes and particular contextual information about the site of data collection (Palmer et al., 2017), is a good example on how to design metadata tools driven by stakeholder needs and aiming at sufficiency (White, 2014).

The aim of this study is to present a comprehensive overview of the use of metadata tools in the Ecology domain through time. A meta-analysis focused on scientific literature on research data management in the field of Ecology was undertaken to support the identification and discussion of major initiatives, challenges and perspectives in research data management in this domain.

2. The Ecology domain

Ecological informatics is an interdisciplinary field that includes conceptual and methodological tools for the understanding, generation, processing and dissemination of various types of ecological data (Michener, Brunt, & Vanderbilt, 2002). Ecological informatics contributes to: (I) Experimental design phase; (II) Data plan; (III) Data acquisition and management; (IV) Quality assurance and control (QA/QC); (V) Metadata implementation; (VI) Data archival; (VII) Data access and dissemination; and (VIII) Data publication (Michener et al., 2002). For phases (I) and (II), designing the structure of datasets and implementing a logical structure within and among datasets can simplify data acquisition, entry, storage, retrieval and manipulation (Michener et al., 2002). In phase (III) the way in which data are acquired also affects data quality by influencing the amount of human error introduced into measurements. Phase (IV) refers to (QA/QC) strategies that are designed to avoid the introduction of errors, or data contamination into a dataset and the metadata in phase (V) is defined as "data about data" (NISO, 2004), so the datasets need to be described in their content, quality, structure, and accessibility (Michener et al., 2002). Different metadata standards have been developed to assure the description of datasets. Some are more generic and domain-neutral, like Dublin Core (Michener et al., 2002; Weibel, Kunze, Lagoze, & Wolf, 1998), while others are tailored to the biodiversity and ecological communities, such as Ecological Metadata Language (EML) (Michener, Brunt, Helly, Kirchner, & Stafford, 1997; Michener et al., 2002), and Darwin Core (Baker, Rycroft, & Smith, 2014). Others, like the EU INSPIRE Directive 2007/EC are specific metadata models, in this case for spatially explicit datasets (da Silva et al., 2014). Phase (VI) Data archival refers to assemblages of datasets packages that are stored, so users can locate, acquire, understand and use the data (Michener et al., 2002). Phase (VII) for data access and dissemination, and (VIII) for data publication ensure overall access to the datasets.

Ecological informatics is thus a framework that enables scientists to generate new knowledge through innovative tools, approaches and solutions that have been developed over the past decade, increasing scientists' efficiency and supporting faster and easier data discovery, integration and analysis; however, many challenges remain, especially in relation to incorporating Ecological informatics practices into mainstream research and education (Michener & Jones, 2012).

Ecological data cover a wide range of topics such as biodiversity surveys, measurements of environmental condition, inventories of species names and synonyms, species distributions, images and sounds, ecological interactions, behaviour, data set descriptions, and analyses and interpretations (Costello, Michener, Gahegan, Zhang, & Bourne, 2013). The variety of the ecological data makes it difficult to create simple, standardized methods to share resulting datasets, and consequently ecological data is currently described using several metadata models (D. Higgins, Berkley, & Jones, 2002) Further, usually data repositories have limited interoperability due to a lack of standards for data and communication protocols (Wieczorek et al., 2012). Inconsistent and ambiguous terminology in the description of biological datasets also creates obstacles in numerous aspects of data integration and use, including discovery, comparison, and quality assessment. It also makes data reuse by other scientists difficult (Baker et al., 2014; Wieczorek et al., 2012).

The need to start collaborative, multi-disciplinary research programs has been highlighted in order to overcome the challenge of efficiently and comprehensively collecting, documenting, communicating, and ultimately preserving primary research data (Jones et al., 2007). In fact, scientists, professional societies and research sponsors are recognizing the value of data as a product of the scientific enterprise and placing increased emphasis on data stewardship, data sharing, openness and supporting study repeatability (Michener & Jones, 2012). Various initiatives (from legal directives to informatics platforms) were developed to enable the sharing of ecological data:

1. Knowledge Network for Biocomplexity (KNB) (Berkley et al., 2009);
2. INSPIRE (Jones et al., 2007);
3. LTER (Michener, Porter, Servilla, & Vanderbilt, 2011);
4. Map of Life (Jetz, McPherson, & Guralnick, 2012);
5. GBIF (Costello et al., 2013).

Data repositories have also been growing rapidly and hold a tremendous promise for increasing the scope, coverage and societal relevance of ecological and biodiversity studies. Nevertheless, the data in these repositories still do not represent a reasonable portion of the massive ecological, environment and biodiversity data that are collected each year (Berkley et al., 2009).

3. Methods

For this review and for the meta-analysis performed, keywords or expressions based in the core area (i.e. Ecology, including Biodiversity), and then specific keywords from research data management (i.e. metadata and data management), were selected. The rationale behind the choice of keywords was to capture as many papers as possible in the Ecology domain and, more specifically, within data management. The Keywords selected were 'Metadata' OR 'Ontology-based approach' OR 'Data management' AND 'Ecolog*' OR 'Biodiversity', then redefined by the following scientific areas: Computer Science Information Systems, Computer Science Theory Methods, Computer Science Interdisciplinary Applications, Computer Science Hardware Architecture, Information Science, Library Science and Computer Science Software Engineering. This was done in order to capture papers within the research data management area ('Biodiversity' and 'Ecology' were not used because it was already in the keywords).

The time span of the search was 1900 to 2018. Searches were carried out between October 2017 and March 2018. ISI Web of Science (ISI WOS; http://webofknowledge.com/) was used, since it offers the widest coverage of published scientific literature (Buchadas et al., 2017; J. P. Higgins & Green, 2011).

However, records gathered from Google Scholar that were absent from the ISI Web of Knowledge search were added to the final dataset. The inclusion criterion was to encompass works in the field of ecology and biodiversity with metadata methods (e.g. metadata models, data repositories, metadata language, data management). The selection was performed by individually examining first the title, keywords and abstract, and then the full text of the scientific manuscript. The exanimation of the papers and the decision on inclusion were made by an expert in the field of ecology and biodiversity.

4. Results and Discussion

The number of records retrieved from ISI Web of Science when using ‘Metadata’ as keyword was 15360. However, when including ‘Ontology-based approach’ as additional keyword, 15981 records were obtained. When including also keywords related to ‘Ecolog*’ OR ‘Biodiversity’ the number of records decreased to 368. After refining per scientific areas, the final number of records was 75 (in October 2017) and 126 (in March 2018) (Table 1).

Table 1 - Number of records retrieved in the literature search in the ISI Web of Science.

Keywords (General)	Keywords (Domain specific)	Results
Metadata		15360
"metadata" OR "Ontology-based approach“		15981
"metadata" OR "Ontology-based approach“ OR “data management”	"Ecolog*"	288
"metadata" OR "Ontology-based approach“ OR “data management”	"Ecolog*" OR "Biodiversity"	368
	redefined by scientific areas	75
"metadata" OR "Ontology-based approach“ OR “data management”	"Ecolog*" OR "Biodiversity"	681
	redefined by scientific areas	126

The final subset of records included in this study was 93 (from the first and second literature search lists, 75 and 126 records, retrieved from ISI WOS) when applied the inclusion criterion (Fig.1).

When analysing the temporal evolution of the use of metadata tools in the scientific domain in the final subset of 93 records (Fig. 1), an increasing number of records in recent years is observed. One possible explanation is the gradual increase of the awareness towards the importance of metadata as a way to improve data management and data repository services.

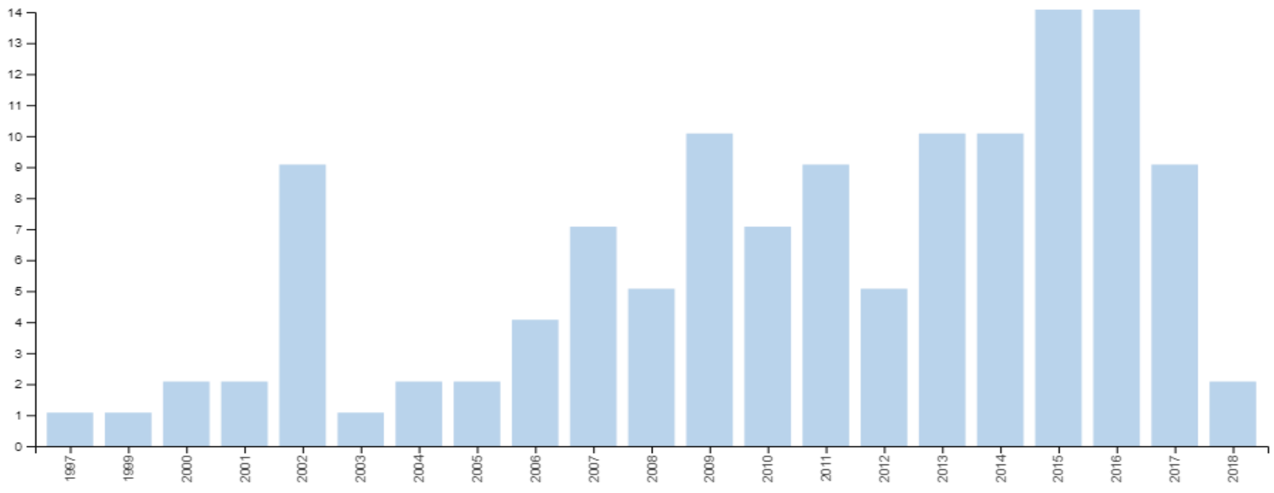


Figure 1 - Number of records retrieved from the literature search in ISI Web of Science per year (temporal overview).

We found 40 different metadata tools either mentioned or used in the scientific manuscripts analysed from 1997 (the data of the first record) to 2018 (see Fig. 2). Each line of the graphic on Figure 2 corresponds to a metadata tool found in more than 3 records, while dots correspond to metadata tools that appear 3 times in the records analysed. The scale in the left refers to the number of records represented by the lines. The scale in right represent the number of records showed by the dots. Metadata tools that only appear once or twice were excluded from the graphical representation in order to simplify the visualization.

Most of these metadata tools are schemas and standards, such as: ABCD schema (Access to Biological Collection Data); FGDC Content Standard for Digital Geospatial Metadata; Crystallographic Information File (CIF); Darwin Core; Data Documentation Initiative; Directory Interchange Format (DIF); Dublin Core; Ecological Metadata Language; GML; Humboldt Core; ISO 19115 and its adoption by the INSPIRE metadata guidelines; ISO 19139; Macromolecular Crystallographic Information File (mmCIF); MIAME Notation in Markup Language (MINiML); Micro-Array Gene Expression Markup Language (MAGE-ML), ThermoML.

Other metadata tools range from metadata catalogues, ontologies, profiles and extensions to metadata editors and encoding standards, namely the: FGDC/CSDGM Biological Data Profile, Darwin Core (semantic web), Encoded Archival Description, Global Change Master Directory´s Interchange Format, iQL, Metacat, MIMOSA/ISO based XML schema, Morpho, NDG models, NEXML ThermoML, OBOE, SEEK, TERN Eco-Informatics data portal known as ÆKOS.

Our results showed that 14 different metadata tools were found to be used in the Ecology domain more than once in the records analysed.

The Ecological Metadata Language (EML) is a metadata standard widely applied in projects and platforms, since its year of implementation, 1997 (Aloisio, Milillo, & Williams, 1999; Michener et al., 1997). INSPIRE is based on the infrastructures for Spatial information established and operated by the European Union and was implemented in 2007. INSPIRE is the first "regional approach" and a legislative attempt to harmonize metadata standards for spatially explicit data (Filetti & Gnauck, 2011). The Darwin Core standard is used for sharing data about biodiversity and it first emerged in 1999 (Wieczorek et al., 2012).

Our review also revealed 50 different platforms/projects in ecology and biodiversity with the specific aim to encourage scientists to share, describe and publish their data. In Table 2, we list examples of such platforms/projects and the associated metadata standard. These examples, to date, are still available in the corresponding website and were mentioned in more than 1 record from the manuscripts analysed. Long-term Ecological Research (LTER) was initiated in 1980 for 6 sites, but this network has been increasing its reach globally since then (Michener et al., 2011). Likewise, the Knowledge Network for Biocomplexity (KNB) data repository has grown fast and now contains over 15,000 datasets (Berkley et al., 2009). The Taxonomic Databases Working Group (TDWG) was created to support the international collaboration of biodiversity informatics institutions and projects, to establish, adopt and promote standards and guidelines for the recording and exchange of data about organisms around the world (Veiga et al., 2017). The global initiative 'Map of Life' aims to gather, store and analyse data from species occurrences, fostering current knowledge on species distribution and contributing to reporting processes (Jetz et al., 2012). The Global Biodiversity Information Facility (GBIF) was created in 1999 and is currently the largest platform with more than seven hundred million occurrence records provided from more than 50 countries (Veiga et al., 2017). Other important initiatives in biodiversity and ecology are IPBES (Intergovernmental Platform on Biodiversity and Ecosystem Services) and GEO BON (Group on Earth Observation Biodiversity Observation Network) (Guralnick, Walls, & Jetz, 2017). Other datasets and data repositories retrieved in this review were: Forest Science Data Bank (FSDB), The Canopy Database Project, The Jalama Project, The Science Environment for Ecological Knowledge (SEEK), The BioCORE Project, The National Biological Information Infrastructure (NBII), Data ONE, The 'BEFdata' platform, BIOFRAG and IRBAS (The Intermittent River Biodiversity Analysis and Synthesis) (Cushing et al., 2007; Gil, Hutchison, Frame, & Palanisamy, 2010; Henshaw, Spycher, & Remillard, 2002; Leigh et al., 2017; Malaverri, Vilar, & Medeiros, 2009; Michener et al., 2007; Nadrowski et al., 2013; Pfeifer et al., 2014; Reichman, Jones, & Schildhauer, 2011).

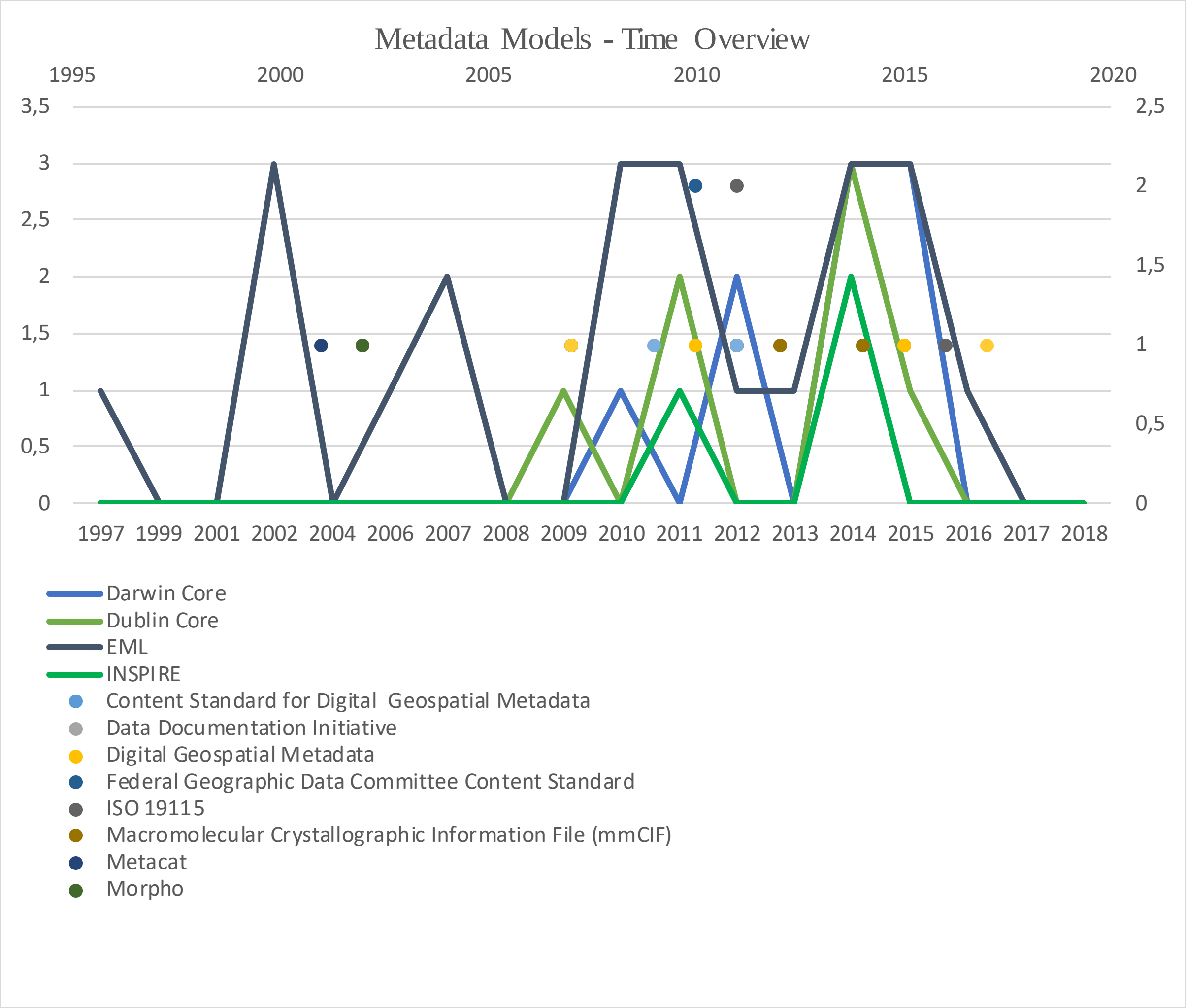


Figure 2 - Metadata tools retrieved from the records analysed through time (1997-2018).

Table 2 - Examples of platforms/initiatives developed and implemented to share, describe and publish data on the fields of ecology and biodiversity, and the associated metadata standards.

Platforms/Projects	Metadata Standard
Long-term Ecological Research (LTER)	EML
The Knowledge Network for Biocomplexity (KNB)	EML
Taxonomic Databases Working Group (TDWG)	Darwin Core
Map of Life Project	Darwin Core
Global Biodiversity Information Facility (GBIF)	Darwin Core

5. Conclusions and Future Perspectives

Since the 1990's the number of metadata tools referred and used by researchers in the field of Ecology has been increasing, alongside with the number of global and national/regional initiatives developed and implemented to share data according to common standards among researchers. With the development of metadata and initiatives to collect, store and share data among researchers, a wide range of metadata tools is currently available to researchers in the field. The 'big data' era further contributes to a pressing need to describe and publish data, so that it can be used within the same research area, as well as across research disciplines.

With an increasing number of initiatives, platforms and repositories that can be used to deposit, publish and share their data with fellow scientists, researchers face new challenges. Such challenges relate e.g. to the lack of comprehensive metadata models that can be used to describe the various types of data used in the domain of Ecology. In many cases, researchers describe available datasets within the context of project consortia, when they are faced with the need to describe the data to be shared with fellow scientists. However, selecting and following a specific metadata model is not an easy task. A major challenge is to guarantee that previous metadata can be harmonized, so that previous work done by researchers is not lost.

Another relevant challenge is the complexity of the available metadata models. In fact, most metadata models available were not developed specifically to describe data in the domain of Ecology. A possible solution is proposed by Qin and Li (2013) consisting in a flexible ontology-based approach to break complex metadata standards into independent modules, so that metadata elements can be optimized for specific needs, while inconsistencies in naming conventions are also addressed. There is, thus, the pressing need to develop interdisciplinary research towards the development of suitable and easy to use metadata models and standards to foster data sharing and publication in the domain of Ecology.

Acknowledgements

This research is a result of the project TAIL – Research data management from creation to deposit and sharing - POCI-01-0145-FEDER-016736, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by National funds through FCT- Portuguese Science Foundation. (PTDC/EEI-ESS/1672/2014/2014).

References

Aloisio, G., Milillo, G., & Williams, R. D. (1999). An XML architecture for high-performance web-based analysis of remote-sensing archives. Future Generation Computer Systems, 16(1), 91-100. doi:10.1016/s0167-739x(99)00038-2

Baker, E., Rycroft, S., & Smith, V. S. (2014). Linking multiple biodiversity informatics platforms with Darwin Core Archives. Biodiversity Data Journal(2), e1039. doi:10.3897/BDJ.2.e1039

Ball, A., Greenberg, J., Jeffery, K., & Koskela, R. (2016). RDA Metadata Standards Directory Working Group.

Berkley, C., Bowers, S., Jones, M. B., Madin, J. S., Schildhauer, M., & Ieee. (2009). Improving Data Discovery in Metadata Repositories through Semantic Search.

Berkley, C., Jones, M., Bojilova, J., & Higgins, D. (2001). Metacat: A schema-independent XML database system.

Buchadas, A., Vaz, A. S., Honrado, J. P., Alagador, D., Bastos, R., Cabral, J. A., . . . Vicente, J. R. (2017). Dynamic models in research and management of biological invasions. Journal of Environmental Management, 196, 594-606. doi:https://doi.org/10.1016/j.jenvman.2017.03.060

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution, 28(8), 454-461. doi:http://dx.doi.org/10.1016/j.tree.2013.05.002

Cushing, J. B., Nadkarni, N., Finch, M., Fiala, A., Murphy-Hill, E., Delcambre, L., & Maier, D. (2007). Component-based end-user database design for ecologists. Journal of Intelligent Information Systems, 29(1), 7-24. doi:10.1007/s10844-006-0028-6

da Silva, J. R., Castro, J. A., Ribeiro, C., Honrado, J., Lomba, A., & Goncalves, J. (2014). Beyond INSPIRE: An Ontology for Biodiversity Metadata Records. In R. Meersman, H. Panetto, A. Mishra, R. ValenciaGarcia, A. L. Soares, I. Ciuciu, F. Ferri, G. Weichhart, T. Moser, M. Bezzi, & H. Chan (Eds.), On the Move to Meaningful Internet Systems: Otm 2014 Workshops (Vol. 8842, pp. 597-607).

European Commission. Guidelines on FAIR Data Management in Horizon 2020. (2016). Retrieved from http://ec.europa.eu/research/%0Aparticipants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Filetti, M., & Gnauck, A. (2011). A Concept of a Virtual Research Environment for Long-Term Ecological Projects with Free and Open Source Software. In J. Hrebicek, G. Schimak, & R. Denzer (Eds.), Environmental Software Systems: Frameworks of Eenvironment (Vol. 359, pp. 235-244).

Gil, I. S., Hutchison, V., Frame, M., & Palanisamy, G. (2010). Metadata Activities in Biology. Journal of Library Metadata, 10(2-3), 99-118. doi:10.1080/19386389.2010.506389

Guralnick, R., Walls, R., & Jetz, W. (2017). Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. Ecography, n/a-n/a. doi:10.1111/ecog.02942

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280-299. doi:10.1353/lib.0.0036

Henshaw, D. L., Spycher, G., & Remillard, S. M. (2002). Transition from a legacy databank to an integrated ecological information system.

Higgins, D., Berkley, C., & Jones, M. B. (2002). Managing heterogeneous ecological data using Morpho.

Higgins, J. P., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions (Vol. 4): John Wiley & Sons.

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology & Evolution, 27(3), 151-159. doi:http://dx.doi.org/10.1016/j.tree.2011.09.007

Jones, C., Blanchette, C., Brooke, M., Harris, J., Jones, M., & Schildhauer, M. (2007). A metadata-driven framework for generating field data entry interfaces in ecology. Ecological Informatics, 2(3), 270-278. doi:http://dx.doi.org/10.1016/j.ecoinf.2007.06.005

Leigh, C., Laporte, B., Bonada, N., Fritz, K., Pella, H., Sauquet, E., . . . Datry, T. (2017). IRBAS: An online database to collate, analyze, and synthesize data on the biodiversity and ecology of intermittent rivers worldwide. Ecology and Evolution, 7(3), 815-823. doi:10.1002/ece3.2679

Lyon, L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report.

Malaverri, J. G., Vilar, B., & Medeiros, C. B. (2009). A TOOL BASED ON WEB SERVICES TO QUERY BIODIVERSITY INFORMATION.

Mayernik, M. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators.

Michener, W. K., Beach, J. H., Jones, M. B., Ludascher, B., Pennington, D. D., Pereira, R. S., . . . Schildhauer, M. (2007). A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems, 29(1), 111-126. doi:10.1007/s10844-006-0034-8

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES. Ecological Applications, 7(1), 330-342. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Michener, W. K., Brunt, J. W., & Vanderbilt, K. L. (2002). Ecological informatics: A long-term ecological research perspective.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supportingecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85-93. doi:http://dx.doi.org/10.1016/j.tree.2011.11.016

Michener, W. K., Porter, J., Servilla, M., & Vanderbilt, K. (2011). Long term ecological research and information management. Ecological Informatics, 6(1), 13-24. doi:http://dx.doi.org/10.1016/j.ecoinf.2010.11.005

Nadrowski, K., Ratcliffe, S., Bönisch, G., Bruelheide, H., Kattge, J., Liu, X., . . . Wirth, C. (2013). Harmonizing, annotating and sharing data in biodiversity–ecosystem functioning research. Methods in Ecology and Evolution, 4(2), 201-205. doi:10.1111/2041-210x.12009

Nilsson, M. (2008). The Singapore framework for Dublin Core application profiles. ttp://dublincore. org/documents/singapore-framework/.

Palmer, C. L., Thomer, A. K., Baker, K. S., Wickett, K. M., Hendrix, C. L., Rodman, A., . . . Fouke, B. W. (2017). Site-based data curation based on hot spring geobiology. PLOS ONE, 12(3), e0172090.

Pfeifer, M., Lefebvre, V., Gardner, T. A., Arroyo-Rodriguez, V., Baeten, L., Banks-Leite, C., . . . Ewers, R. M. (2014). BIOFRAG – a new database for analyzing BIOdiversity responses to forest FRAGmentation. Ecology and Evolution, 4(9), 1524-1537. doi:10.1002/ece3.1036

Qin, J., & Li, K. (2013). How portable are the metadata standards for scientific data? a proposal for a metadata infrastructure. Paper presented at the International Conference on Dublin Core and Metadata Applications.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. Science, 331(6018), 703-705. doi:10.1126/science.1197962

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. Information Processing & Management, 49(6), 1194-1205.

Thanos, C. (2017). Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies. Publications, 5(1), 2.

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE, 12(6), e0178731. doi:10.1371/journal.pone.0178731

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (2070-1721).

White, H. C. (2014). Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors. Journal of Library Metadata, 14(1), 24-51. doi:10.1080/19386389.2014.891896

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., . . . Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE, 7(1), e29715. doi:10.1371/journal.pone.0029715

Metadata Models for Organizing Digital Archives on the Web: Metadata-Centric Projects at Tsukuba and Lessons Learned

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

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

Chiranthi Wijesundara
University of Tsukuba,
Japan
chiranthis@gmail.com

Winda Monika
University of Tsukuba,
Japan
windabi.wm@gmail.com

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

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

Abstract

There exist many digital collections of cultural and historical resources, referred to as digital archives in this paper. Domains of digital archives are expanding from traditional cultural heritage objects to new areas such as popular culture and intangible entities. Though it is known that metadata models and authority records, such as subject vocabularies, are essential in building digital archives, they are not yet well established in these new domains. Another crucial issue is semantic linking among resources within a digital archive and across digital archives. Metadata aggregation is an essential aspect for the resource linking. This paper overviews three metadata-centric on-going research projects by the authors and discusses some lessons learned from them. The subject domains of these research projects are disaster records of the 2011 Great East Japan Earthquake, Japanese popular culture such as Manga, Anime and Games, and cultural heritage resources in South and Southeast Asia. The main goal of this paper is not to report on these projects as completed research, but to discuss issues of metadata models and aggregation which are important in organizing digital archives in the web-based information environment.

Keywords: digital archives; digital collections; digital curation; digital libraries; fan-created websites; Japanese popular culture (pop-culture); manga; linked open data; memory institutions; metadata aggregation; tangible and intangible cultural heritage

1. Introduction

Since the early 1990s, many digital collections of cultural, historical, and scholarly resources have been developed by many memory institutions, such as libraries, museums, and archives. Those collections were built to improve user accessibility to important cultural resources via the Internet while simultaneously preserving those resources. It is widely known that metadata is key to building those digital collections for both preservation and access, as metadata is needed to handle digital resources as well as cultural heritage objects in various aspects, i.e., search, select, organize, access, and preserve. Meanwhile, major activities to develop metadata standards oriented to the web environment also started in 1990s, e.g., Dublin Core Metadata Initiative (DCMI, http://dublincore.org/), Open Archival Initiative (OAI, http://www.openarchives.org/), and so on.

Metadata aggregation from multiple sources has been widely recognized as an important technology to create and organize a merged digital collection on the web. OAI developed a metadata harvesting protocol named OAI-PMH which is used to build value-added services. For example, National Science Digital Library (NSDL, https://nsdl.oercommons.org) collected and aggregated metadata from many sources. Europeana and Digital Public Library of America (DPLA), well known as large portals for digital collections of cultural resources in Europe and the United States, respectively, harvest metadata from their participating institutions. In Japan, the

National Diet Library (NDL) has built a portal which collects metadata using OAI-PMH and provides unified access to the digital collections of Great East Japan Earthquake developed by regional public sectors, NPOs, universities, news and information companies, etc. Thus, metadata harvesting is a commonly used technology to add values to digital collections of cultural resources. In the current web environment, Linked Open Data (LOD) technologies provide information environments to develop more sophisticated services using various LOD datasets in addition to those digital collections.

Digital collections and services are sometimes referred to by different names, e.g., digital library, digital museum, digital archive and digital gallery. We use the term *digital archive* to refer to those digital collections and services for the rest of this paper. This is because the term *archive* has the meaning to collect important resources to provide access over time for future use. We call digital archives created by memory institutions using their holdings *institutional digital archives*.

Many digital archives developed by memory institutions provide digital copies of cultural heritage objects held by the institutions. They are built using the catalog data of the institution as the base metadata for the digital archives. In most cases of existing institutional digital archives, those digital copies are created by digitizing the original cultural heritage objects using devices such as digital cameras, scanners, analog-to-digital converters, etc. Even in cases where institutions cannot allow open access to the digitized copies of the objects, the metadata of the cultural objects plays important roles for users in finding and accessing the digital resources and the original cultural objects.

In general, while institutional metadata developed by memory institutions to describe their holdings is highly standardized, it often provides very limited information. A typical problem is that they do not provide contextual information of cultural heritage objects which would help users understand their values. On the other hand, websites such as Wikipedia and those created by fans and specialists provide in-depth descriptions about the objects, including contextual information of the objects, but they are less standardized. Aggregation of different types of cultural heritage object descriptions, which we call metadata aggregation, is important because we can obtain better description of cultural heritage objects by combining institutional metadata and other websites. Metadata in domains such as pop-culture and intangible cultural heritage are not well standardized like those of traditional cultural heritage objects. Metadata aggregation is very important in these domains. This viewpoint is shared among the three projects shown in this paper.

This paper describes three research projects that have been undertaken to address these issues. While not intended as a report of completed research projects, this paper seeks:

- to discuss challenges in metadata aggregation in domains which are not well covered by conventional institutional digital archives in the cultural domain,
- to discuss the development of underlying models of metadata and some lessons learned from the research in the three domains – disaster records, Japanese pop-culture and intangible cultural heritage, and
- to re-think a few general metadata models such as One-to-One Principle (Miller, 2010; Woody, Clement & Winn, 2005), FRBR (IFLA, 2009) and Metadata Application Profiles (Nilsson, Baker & Johnston, 2008).

The rest of this paper is organized as follows. In Section 2, we discuss metadata aggregation as a key issue to enhance usability of the digital archives. Section 3 shows the three projects followed by discussions on lessons learned from the projects and conclusion in Sections 4 and 5, respectively.

2. Backgrounds

2.1. Metadata Aggregation for Digital Archives of Cultural and Historical Resources – Basic Concepts and Issues

Metadata aggregation is broadly used to collect metadata from multiple repositories and organize a virtually unified repository as mentioned earlier. OAI's metadata harvesting protocol (OAI-PMH) is a widely used protocol for collecting metadata.

Europeana Data Model (EDM) (Isaac, 2013) defines entities expressed in their metadata such as digital images, original cultural heritage objects, and relationships among those entities. EDM defines aggregation both within a single digital archive and across digital archives.

Ministry of Internal Affairs and Communication (MIC) of the Japanese government supported regional public sectors to digitally archive the records of the Great East Japan Earthquake occurred in March 2011. NDL worked as a national institution to help organize the regional digital archives and built a national portal for the disaster archives, which is named Hinagiku (http://kn.ndl.go.jp). Hinagiku collects metadata from participating archives. A crucial issue of Hinagiku is the quality of the metadata. The quality of the metadata of the participating archives are, in general, not high mainly because of the limitation of financial, human and time resources; the majority of the resources are photographs and the metadata was created by third parties who were not metadata specialists and within a very limited time. Another aspect is lack of standards to organize metadata, e.g., granularity of the archived resources.

Agency for Cultural Affairs (ACA, Bunka-cho in Japanese) has been hosting the Japan Media Arts Festival for since 1997, which covers four resource types: Art, Entertainment, Animation and Manga. The Media Arts Database (MADB, https://mediaarts-db.bunka.go.jp) hosted by ACA collects metadata in the four resource types Manga, Animation, Game and Media Art. (notes: Manga is a Japanese term meaning Japanese comics or graphic novels. Animation is written as Anime in this paper.)

Anime and Games are typically part of large franchises, e.g. *Dragon Ball, Gundam, Pokémon*, and so on. Metadata aggregation for members of these franchises from different mediums is a crucial requirement to enhance the usability of the MADB. Linking contents of different media types by shared franchise can improve features such as search and retrieval. To achieve this, however, we need to model the franchise as an entity like the Work and Item entities of FRBR, and relationships among franchise entities and other entities described by MADB.

2.2. Data Model Issues for Metadata Aggregation

Data models have crucial roles for defining metadata schemas interoperable with other schemas. EDM and OAI-ORE define structures of metadata collected and aggregated from multiple metadata sources. Metadata mapping is a key issue in the aggregation process in the case that metadata to be aggregated are created on different schemas. Metadata mapping is often done only based on properties. However, property-level mapping has risks of losing context of properties given in the schema in which the properties are included, such as mandatory levels and value types. Thus, we need to use contextual information for metadata mapping.

FRBR provides us with generalized object types for books and other published materials. MADB in part adopts FRBR to define the object types for their databases. In our project on Japanese pop-culture, we defined our data models based on those object types of MADB. In particular for Manga, we defined a three-layer model composed of *Superwork*, *Work* and *Volume*.

One-to-One principle of metadata is well-known as a basic model which says that relationships between metadata and its objective resource should be One-to-One. A typical One-to-One metadata example is a catalog record of a physical object at a memory institution. Non-One-to-One metadata is not uncommon; typical digital archive metadata has descriptions about the original cultural heritage objects, descriptions about a digital resource(s) created from the original object and some other related descriptions in the digital archives hosted by the memory institutions. This model which has an original cultural heritage object fits well to tangible cultural heritage objects.

However, it does not work well for those cultural heritage which cannot be collected by memory institutions as a physical item, e.g., intangible cultural heritage (e.g., performance and craftsmanship), very large objects (e.g., monuments and large statues) and any objects which have short life-time (e.g., installation arts, ice arts and fireworks). A typical solution for this kind of cultural heritage is to use photographs or videos as a surrogate of the heritage object. In this case, we would need description not only about an original cultural heritage object but also about its surrogate. Identification of the cultural heritage objects as an objective of metadata description is crucial if the relationship between the heritage objects and their metadata is non-One-to-One.

In the case of the disaster archives of Great East Japan Earthquake, their metadata model is a simple One-to-One, i.e., one metadata record for one resource such as a document, photograph, video, or oral record. This simple model works well for metadata creation. However, it often brings us too many hits for a simple search query. The records collected were primarily about an event or a thing related to the disaster, e.g., photographs of damaged local harbors, photographs taken at local festivals, news documents issued by local governments, and so forth. So, it is a natural demand to make groups of the items based on the events, but metadata of a record item does not include description about the contexts of the item which is meaningful for end-users and useful for the grouping of those items. However, identifying those events was not an easy task for the third-party catalogers and metadata description scheme for those events was not well defined.

2.3. Building Digital Archives of Non- Conventional Domains: Research Questions

The access environments for general users to obtain cultural information on the web has been increasingly developed since 1990s, i.e., directory services, full text search engines, blogs, and SNS. Another significant factor of cultural information on the web is the development of cultural resources by non-memory institutions and by crowds. The value of institutional digital archives and their portals are well-recognized. On the other hand, quite a lot of cultural information resources are available in those non-institutional resources such as Wikipedia and those sites created by domain specialists and fans. Semantic Web and Linked Open Data technologies obviously have crucial roles for linking the resources across the boundaries between institutional and non-institutional resources.

There exist digital archives of various types of cultural objects and the boundaries of the archives are becoming fuzzier; from traditional tangible cultural heritage objects to intangible cultural heritage objects such as dance, festivals, cuisines, and craftsmanship; from printed books to electronic and web-based books; from traditional arts to media arts such as videogames and computer arts; from a movable objects to immovable objects such as large statues and buildingss; from simple photographic images to 3D and virtual reality images, and motion graphics.

These digital archives create metadata designed in accordance with their archived resources. Conventional institutional digital archives use metadata standards used by memory institutions as the basis for the digital archives. However, the metadata in the new domains need some new features which are not covered by metadata of those conventional digital archives. For example, in the case of Manga, metadata schemas specialized for describing Manga are not well developed.

Web resources such as Wikipedia, fan-created sites, and exhibition pages by museums often provide richer information for users than institutional metadata which is primarily bibliographic descriptions of their holdings. On the other hand, the institutional metadata provides authoritative information. Therefore, metadata aggregation across institutional digital archives and websites is a crucial issue to build better environment for users on the web to find and access digital resources of cultural domains. The challenge is to bridge the gap between the metadata of these different types of resources, i.e., well-structured institutional metadata and non- or semi-structured descriptions in webpages.

3. Research Projects on Metadata for Digital Archives at Tsukuba

This section briefly describes three research projects at the authors laboratory from which we have learned issues on metadata to improve usability of digital archives.

3.1. Enhancing Usability of Digital Archives of Great East Japan Earthquake by Metadata Aggregation within and across Archives

(1) Project Background

As mentioned earlier in this paper, there are many digital archives of the Great East Japan Earthquake on March 11, 2011. NDL is running a portal named Hinagiku for those archives. Each of the archives collects various types of resources. Photographs taken by digital cameras are the largest portion of the archived resources. This feature is a significant difference from the archive for the earthquake in January 1995 in Kobe/Awaji because the major archived resources was still paper centric. The Japanese government defined a guideline to develop disaster archives in which they suggested a simple metadata scheme defined based on Simple Dublin Core[1] because, on one hand, the archive resources had to be collected and organized as a database in limited time and costs and, on the other hand, the metadata should be interoperable across digital archives.

We analyzed metadata from five disaster archives that participate in Hinagiku. As these archives were developed by different sectors, each archive had their own features. A feature common among the archives was that a metadata was basically created for every single item and, in general, the quality of those metadata was not high as mentioned earlier. As a result, they had common usability issues; for example, numerous results for a simple query, such as multiple photographs taken at a single location, and low-quality descriptions in the subject and title fields of metadata.

(2) Research Problems

A fundamental problem we have found is the need to create a set of items by aggregating metadata and providing users with set-level access functions in addition to item-level access. For example, photographs taken at a single event should be aggregated as a set of photographs of the event. We need to automatically create metadata for an aggregated instance. However, this aggregation is not a simple task because contextual information to semantically group the items is not given in most cases.

(3) Approaches and Some Findings

The followings are approaches in our project,

(a) metadata aggregation
- aggregating metadata of photographs by time and location (longitude/latitude) information
- aggregating metadata by co-occurrence of subject terms

(b) building regional ontology resources
- domain ontology for local service and sectors using terms extracted from metadata
- datasets to record provenance of geographic entity names such as villages and towns

In the approach (a), we did metadata aggregation by time-location information of photographs collected by the three archives, Aomori Archive, Kuji-Noda-Fudai (KNF) Archive and Michinoku Shinrokuden. Aomori and KNF are developed by regional governments and Shinrokuden is developed by Tohoku University and covers broader area than the former archives. The aggregation process: (1) collect photographs whose location information is within an area represented as a hexagon whose edge length is approximately 1600m, and then (2) sort photographs by time of creation of each of the collected photograph sets and split the photographs if an interval between two consecutive photographs are larger than 30 minutes. For subject-term based aggregation, we applied Levenshtein distance for clustering subject terms. For each term cluster, we aggregated metadata which included at least one subject term contained in the cluster to form an aggregated metadata set. Table 1 and 2 show the results of time-location and subject-terms approaches, respectively (Seki, 2018). The number of resources shown in the tables differs as not all metadata had time-location information or proper subject terms. This results show that we can aggregate resources fairly effectively, but neither of these approaches is perfect. Precise evaluation is still left for our future work.

[1] The metadata schema is defined based on DC-NDL, which is a metadata schema defined by NDL based on Simple Dublin Core.

Table 1 Aggregation by Time and Location

Archives	Metadata	Sets	Size=1	Size>99
Aomori	48,338	6,571	1,599	102
KNF	72,894	8,531	2,023	189
Shinrokuden	96,441	8,188	962	59

Table 2 Aggregation by Subject Terms

Archives	Metadata	Sets	Size=1	Size>99
Aomori	68,032	3,596	1,060	1,332
KNF	127,383	8,290	2,240	609
Shinrokuden	124,552	7,910	2,047	4,612

Metadata: Number of Metadata Instances
Sets: Number of Sets created by Aggregation
Size: Number of Aggregated Sets of the given Size

We have not obtained aggregation results in approach (b) but we have learned from approach (a) that vocabularies to represent regional entities such as place names and organizations are crucial to link data within and across archives. In addition, we are currently developing a dataset which stores change history of place names, e.g. towns and villages, to link data in the disaster archives and some other resources such as Tsunami Digital Library (http://tsunami-dl.jp/) which collects many papers, reports and news articles about disasters caused by Tsunamis in Japan since the 1890s.

From our analysis on the metadata of the component archives, we have learned that automated aggregation of contents in a single archive and metadata creation for the aggregated contents is crucial. As each participating archive contains items related to the disaster and events directly or indirectly caused by an earthquake, a crucial issue to improve usability, in particular for the local communities, is to link the disaster archive to other resources in order to cover longer period of time and to help regional communities keep their memory safe and see their history in future.

3.2. Metadata Model for Aggregating Manga Resources

(1) Project Background

Manga, Anime and Video Games are very popular Japanese pop-culture types, each with numerous commercially published materials. However, there is no commonly accepted metadata standards defined for those materials. For example, libraries use MARC for bibliographic descriptions of items of these resource types, while MADB defines its own metadata schema and its underlying data model for each.

Interoperability across the metadata databases of different types was a crucial requirement for the database design of MADB. FRBR was used as an underlying framework in the design of the data models which define classes to represent component instances such as a monograph, a series of monographs, and a game package. Each class is given a set of properties to describe attributes of a component instance. The first author of this paper was involved in the data model design as an advisory group member with some of his co-authors. From the discussions on the data models, the authors learned many interesting features of Manga and other pop-culture resources, which were used in the research project described in the paragraph below.

(2) Research Problems

The primary research issue of the Manga metadata project described in this paper is to develop a metadata model for aggregating Manga metadata of MADB as an institutional metadata and metadata presented in web resources such as Wikipedia and fan-created sites e.g., AniDB and MyAnimeList. The fundamental problem for aggregation is the differences of their description levels; MADB is primarily developed based on Item/Manifestation and those web resources are Work/Expression based. This is quite natural because MADB is developed using bibliographic data of Items held by memory institutions such as NDL and Kyoto International Manga Museum and descriptions in the web resources are likely on Work and Expression levels. The research issue is how to bridge the gap. In our earlier study, we applied EDM for aggregation (Kiryakos & Sugimoto, 2015). In the current research, we are using OAI-ORE as the base aggregation model and have defined a hierarchical model to describe entities of Manga, which may be extended to other resource types (Kiryakos & Sugimoto, 2018). The hierarchical model has three levels – Superwork,

Work and Volume. Work and Volumes roughly correspond to Work/ Expression and Manifestation/ Item (see FIG. 1.). Superwork is defined as an entity aggregating Works in different media created under a single franchise such as Gundam, Dragon Ball, and One Piece (Kiryakos et al., 2017) (Lee et al., 2018).

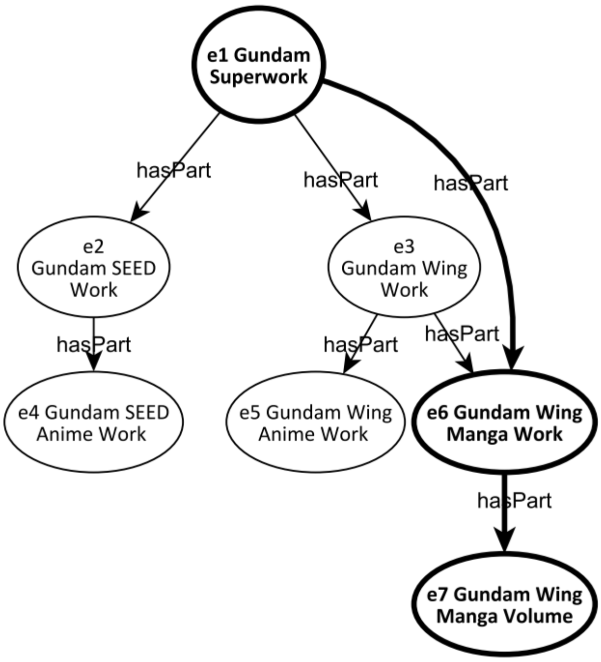


FIG.1. Three Layer Model (Superwork / Work / Volume)

(3) Approaches and Some Findings

Metadata mapping of MADB across resource types: From the advisory discussion for MADB, we have learned that class-level matching using the data models prior to property level mapping across resource types works well for the metadata mapping task. We chose a minimum set of classes defined in each resource type of MADB, and found that identifying corresponding core classes was a key issue in the matching across resource types. These classes may stand for Description Template of the DCMI application profiles.

Metadata aggregation for Manga and other resource types: The focus of our research is to define a simple data model for metadata aggregation using MADB and fan-created web resources. We found FRBR Group 1 entities were useful in their hierarchical structure, but few web resources described entities in the way FRBR had defined them, thus leading to the definition of a model composed of the three levels – Superwork, Work and Volume. We defined Superwork as an entity different from Complex Work of FRBR object oriented (FRBRoo). Note that a detailed discussion about the Superwork entity is not included in this paper.

3.3. Modeling Cultural Heritage Objects for Digital Archives

(1) Project Background

This on-going study involves the creation of a data model to describe cultural heritage objects and their metadata for digital archives with a focus on intangible cultural heritage objects. This project started from the viewpoint of digital archives for South and Southeast Asia (Wijesundara, Sugimoto, Narayan, & Tuamsuk, 2016) (Wijesundara, Sugimoto, & Narayan, 2015). Digital archives by memory institutions in the region are not well developed. However, on the web, we can find rich digital cultural heritage resources of the region developed by memory institutions in Europe and North America. Websites such as Wikipedia also provide detailed description on cultural heritage. Therefore, we think that aggregation of metadata taken from those sites is an important aspect to develop digital archives of cultural heritage in the regions.

(2) Research Problems

Institutional digital archives of cultural heritage are built on institutional metadata. A significant regional issue is the development levels of institutional metadata. A crucial domain-specific issue is the lack of well-developed and standardized metadata models for intangible cultural heritage. As mentioned above, we need to aggregate web resources and the institutional metadata to obtain detailed descriptions. This is the same issue discussed in the previous section on pop-culture metadata. Moreover, objectives of metadata description are not clear in the case of intangible cultural heritage because we can only archive records of a particular performance(s) inherited by a person or a community, e.g., dance, music, craftsmanship, etc. Those recorded materials may be a photograph, video, data captured by sensors, or documents. Thus, collecting performance records is essential to develop digital archives of intangible cultural heritage.

However, we need to archive many tangible objects such as costumes, instruments, and music scores, together with those records to preserve intangible cultural heritage. Moreover, we need to collect descriptions about the cultural heritage to link these tangible objects and recorded materials. "We need to collect information and aggregate them to answer these requirements" was our primary research question. Then, we had further questions "What data model is suitable for intangible cultural heritage?" and "Can we define digital archiving as an intellectual creation activity?"

(3) Approaches and Some Findings

We have defined a model named CHDE (Cultural Heritage in Digital Environment) to describe a process to organize digital archives for both tangible and intangible cultural heritage and to identify entities to be described in metadata about cultural heritage objects and their digital surrogates (Wijesundara, Monika, & Sugimoto, 2017). In the design process of CHDE (see FIG 2.), we applied the One-to-One Principle of Metadata to clearly identify the relationships between metadata and its objective of description in order to avoid confusion in the process of metadata aggregation. We analyzed several cultural heritage metadata taken from institutional and non-institutional data resources, e.g., British Museum and Wikipedia, to define a metadata aggregation scheme based on CHDE. We first categorized the properties of the metadata into four categories, which are original cultural heritage objects, their digital surrogates, administrative information and other miscellaneous entities named as external resources. Then, we grouped the properties in each category into sub-categories based on the classes of description objectives, e.g. agent, location, rights, and so on. The first level categories are useful to identify objectives in non-One-to-One situation, and the second level categories are useful to define metadata mappings across archives.

We proposed a model which uses FRBR Group 1 entities (FRBR WEMI) to identify intellectual creation by curators who organized digital archives of cultural heritage objects (Monika, Wijesundara, & Sugimoto, 2017). We are investigating the applicability of FRBR WEMI to digital curation and exhibition as intellectual creative activities and products by curators. Works of media arts are often dynamic objects, which have similar feature with intangible cultural heritage. The metadata model proposed in this project which is based on CHDE may be applied to those dynamic objects. And, exhibitions are intangible by nature. In general, metadata about exhibition is not a collection metadata because exhibitions are events as well as collections of cultural objects. The metadata model for exhibition will help find and re-use the intellectual creations by curators.

4. Discussions and Lessons Learned

This section discusses some lessons learned from the projects described above.

(1) Data Model Issues for Metadata Aggregation

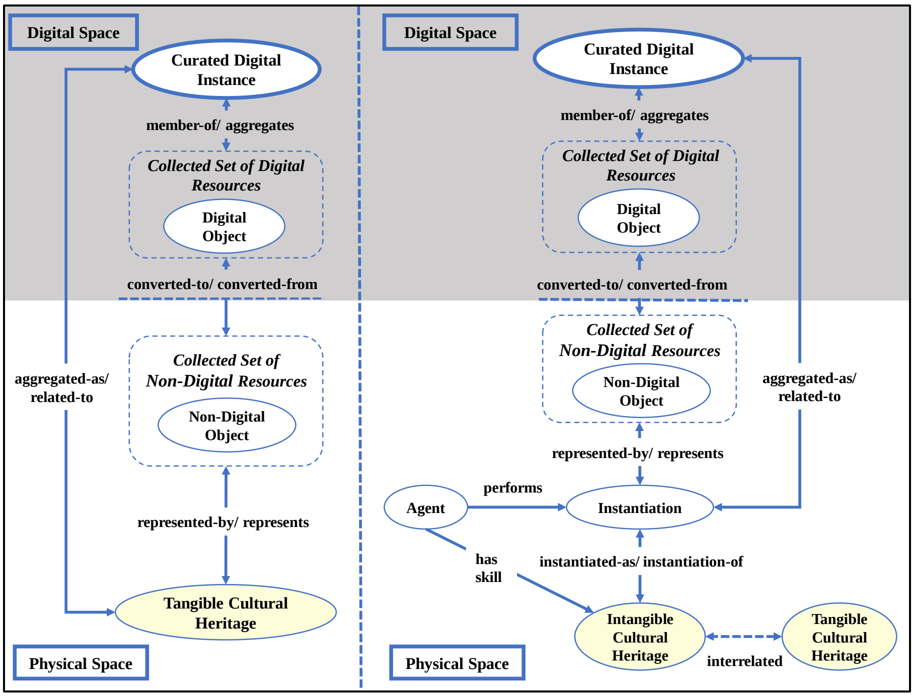

Digital Space
Curated Digital Instance
member-of/ aggregates
Collected Set of Digital Resources
Digital Object
converted-to/ converted-from
aggregated-as/ related-to
Collected Set of Non-Digital Resources
Non-Digital Object
represented-by/ represents
Tangible Cultural Heritage
Physical Space
Digital Space
Curated Digital Instance
member-of/ aggregates
Collected Set of Digital Resources
Digital Object
converted-to/ converted-from
Collected Set of Non-Digital Resources
Non-Digital Object
aggregated-as/ related-to
represented-by/ represents
Agent
performs
Instantiation
has skill
instantiated-as/ instantiation-of
Intangible Cultural Heritage
interrelated
Tangible Cultural Heritage
Physical Space

FIG.2. CHDE Model

Metadata aggregation is a key technology to improve usability of various cultural resources available on the web. There are several different types of metadata aggregation; for example, (A) metadata aggregation based on a standardized protocol among digital archives, (B) metadata aggregation among heterogeneous data sources such as institutional digital archives and websites, and (C) metadata aggregation within a single digital archive. Europeana is type A, the Japanese pop-culture project and cultural heritage project mentioned above are type B, and the metadata aggregation of the disaster archives and analysis shown in section 3.1 is type C. We have learned that the most fundamental issue common among these types is that objectives of metadata description should be identifiable because those objectives are often used as a key for aggregation. This issue is crucial in particular for aggregating metadata built on different schemes (type B). CHDE and the hierarchical model presented above are designed based on this aspect.

(2) Data Model for Digital Archives of Non-Conventional Objects

Conventional digital archives are mostly a collection of digital copies of cultural and historical resources. Those digital copies are surrogates of the original cultural heritage objects which have to be identifiable. Original cultural heritage objects may be easily identified in the case of tangible cultural heritage objects maintained by memory institutions. However, the same scheme may not work well in the case of intangible cultural heritage objects, natural and man-made events, and dynamic objects whose actions have values. In these cases, we need to archive the records of the original objects as their digital surrogates. The relationships among the original objects, their records and the digital surrogates of the records have to be clearly described. Thus, the underlying model for non-conventional digital archives and the model for conventional cultural digital archives are different. This difference may not be significant when developing a single digital archive in a single domain, though it becomes problematic when attempting to link multiple archives and websites. We have learned that we can use One-to-One Principle of Metadata as a simple and clear guideline to define data models to help linking objects across archives and to help clearly identify rights in accordance with each object. There are some arguments on One-to-One but we think it is a crucial concept for digital archives (Urban, 2014).

(3) Requirements for Domain Knowledge – a KOS-oriented View

A common issue for unconventional areas is the lack of controlled vocabularies for describing subjects, types and classes of resources in each area. We are lacking authority records in the non-conventional domain, e.g., descriptive subject headings for pop-culture resources, authority data of characters appear in Manga, Anime and Game, and local terms used in a certain region for disaster archives, and so forth; there is no easy answer to solve this issue. As Wikipedia and other websites are good sources of terms used by fans in the case of Manga and other pop-culture domains, an obvious task would be the extraction of terms from these resources. Automatically identifying semantic relationships between terms, however, would be difficult. Changes to the terms and vocabularies over time is also a crucial issue from the viewpoint of longevity of digital archives.

(4) Metadata Mapping across Different Resource Types

Property level mapping is frequently done for defining a mapping function between different metadata tables. This methodology works well in cases where the tables are not very different, or the tables have fewer numbers of attributes. In the MADB discussion, we have learned that this mapping scheme does not work. We found that we need to identify the entities in each resource type which are the objectives of metadata description, then find entity pairs for mapping across resource types prior to any property level mapping. We need underlying data models explaining entities of the resource types while also defining the classes of those entities. A FRBR-based model was used by the MADB project to identify entities and define their classes for mapping across the three resource types. We think that this class-based mapping framework, which is primarily similar to Description Template-based mapping across different schemas, could be used with various applications that require metadata schema mappings.

(5) Viewing Cultural Digital Archives as Intellectual Creation by Digital Curators

Digital curators who collect digital resources and organize them into a digital archive create various descriptions about the resources and archives, e.g., an exhibition webpage and catalogue. Those descriptions add significant value to the digital resources because they provide contextual information of the cultural resources. An exhibition program may be entirely or partly reused in other events as an intellectual Work. Organization of contents and their visual presentation in an exhibition are Expression of a Work. Digital curators add value to the original cultural objects in various aspects, e.g., selecting, organizing, describing, etc. Clear identification of their intellectual contributions in the organization of digital archive metadata is useful not only to reuse their products but also to create new intellectual products.

5. Concluding Remarks

The research goals of these three projects are not the same and the objectives and content of metadata descriptions are different. However we have discovered problems common across these projects, namely the importance of metadata aggregation for better usability of digital resources, demands to vocabularies based on the domain knowledge, and the importance of underlying data models to connect resources across digital archives of different resource types. We used general models for metadata such as One-to-One Principle and Application Profiles as well as domain specific models such as FRBR and EDM. While these models do not always work well with real-world metadata because of the metadata itself and its schema qualities, we have found that these models have nevertheless played crucial roles in our projects and have taught us valuable lessons.

Acknowledgements

We thank Yusuke Seki and all students who contributed in the research projects presented here. This study is supported by JSPS KAKENHI Grant Number 16H01754.

References

IFLA. (2009). *Functional Requirements for Bibliographic Records*, p. 136, Retrieved, May 2, 2018, from https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

Isaac, A. (ed.). (2013). *Europeana Data Model Primer*. Retrieved, May 14, 2018, from http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

Kiryakos, S., & Sugimoto, S. (2018). Aggregating manga metadata from diverse data providers using the portrayal of bibliographic hierarchies on the Web based on OAI-ORE, Manuscripts in preparation

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

Kiryakos, S., Sugimoto, S., Lee, J. H., Jett, J., Cheng, Y., & Downie, J. S. (2017). Towards a Conceptual Framework for Superworks. *Proceedings of JADH 2017*, pp.47-49.

Lee, J. H., Jett, J., Cho, H.R., Windleharth, T., Kiryakos, S., Disher, T., & Sugimoto, S. (2018). Reconceptualizing Superwork for Improved Access to Popular Cultural, Objects. *Proceedings of ASIS&T 2018*. Manuscript submitted for publication.

Miller, S. (2010). The One-To-One Principle: Challenges in Current Practice. *International Conference on Dublin Core and Metadata Applications*, pp. 150-164.

Monika, W., Wijesundara, C., & Sugimoto, S. (2017). Modeling Digital Archives of Intangible Cultural Heritage Based on One-to-One Principle of Metadata. *Proceedings of the 8th Asia-Pacific Conference on Library and Information Education and Practice (A-LIEP)*, pp. 137-148.

Nilsson, M., Baker, T., Johnston, P. (2008). *The Singapore Framework of Dublin Core Application Profiles*, Retrieved, May 2, 2018, from http://dublincore.org/documents/singapore-framework/

Seki Y. (2018). A Study on Improvement of Usability of Digital Archives of Great East Japan Earthquake using Resource Aggregation, Master Thesis, University of Tsukuba, 46p.

Urban, R. J. (2014). The 1:1 Principle in the Age of Linked Data, *International Conference on Dublin Core and Metadata Applications,* pp.119-128.

Wijesundara, C., Monika, W., & Sugimoto, S. (2017). A Metadata Model to Organize Cultural Heritage Resources in Heterogeneous Information Environments. Choemprayong S., Crestani F., Cunningham S. (eds). Digital Libraries: Data, Information, and Knowledge for Digital Lives. ICADL 2017. *Lecture Notes in Computer Science*, Vol 10647. pp. 81-94, Springer. doi:10.1007/978-3-319-70232-2_7

Wijesundara, C., Sugimoto, S., & Narayan, B. (2015). Documenting Spatial and Temporal Information for Heritage Preservation: A Case Study of Sri Lanka. *Proceedings from the Document Academy*, 2(1), 5.

Wijesundara, C., Sugimoto, S., Narayan, B., & Tuamsuk, K.: Bringing Cultural Heritage Information from Developing Regions to the Global Information Space as Linked Open Data: An Exploratory Metadata Aggregation Model for Sri Lankan Heritage and its Extension. *7th Asia-Pacific Conference on Library and Information Education and Practice (A-LIEP),* pp. 117–132 (2016).

Woodley, M. S., Clement, G., & Winn, P. (2005). *DCMI glossary*. Retrieved, May 2, 2018, from http://dublincore.org/documents/usageguide/glossary.shtml

POSTERS

Linked Data Publishing and Ontology in Korea Libraries
Poster

Mihwa Lee
Kongju National University, South Korea
leemh@kongju.ac.kr

Yoonkyung Choi
National Library of Korea, South Korea
yoonkchoi@korea.kr

Keywords: Linked Open Data; DCTERMS; Dublin Core; metadata vocabularies; ontology; class; property; BIBO; FOAF; SKOS; BIBFRAME

Abstract

This poster was to analyze three linked open data (LOD) services in Korea in an aspect of ontology, and to suggest three LOD to transform their local ontology to BIBFRAME as a measure for interoperability of LOD. For this study, literature review and case studies were conducted. For case studies, KERIS, NLK, and KISTI were selected, which are the major organizations publishing LOD. They have been publishing LOD from bibliographic records and authority data with linking the external LOD such as VIAF, LDS, BNB, ISNI, WorldCat, and so on. We analyzed the characteristics of three LOD according to the following categories: (1) subject domain, (2) volumes of bibliographic, authority, and subject data, (3) ontology, (4) local ontology, and (5) linking external LODs. In particular, in the aspect of ontology, FOAF, SKOS, DC, and BIBO were used in common, and however, MODS, DCTERMS, BIBFRAME, PRISM, and Bibtex were also used in three LOD. Also, three LOD devised their own ontology – properties and classes – due to lack of classes and properties in describing LOD. These local properties and classes were different with inconsistency that would bring out conflicts in data sharing. In an aspect of requirements for metadata, interoperability is very important. Therefore, this study suggested transforming the local ontology of three LOD to BIBFRAME for interoperability and crosswalking.

LOD publishing in Korea

Linked open data (LOD) has been mandatory to construct the semantic web library. In Korea, 10 organizations in a public sector have started their LOD services since 2013. Among them, three organizations provided library-centric LOD, which are KERIS (Korea Education and Research Information Service), NLK (National Library of Korea) and KISTI (Korea Institute of Science and Technology Information).

KERIS has been publishing KERIS LOD (http://data.riss.kr/serviceHome.do) of bibliographic records in late 2013. It has used the properties and the classes from DC, BIBO, MODS, FOAF, SKOS, and KERIS devised local ontology. OCLC WorldCat, LCSH, BNB, GeoNames, DBpedia, Flickr are most consumed for linking vocabularies of external LOD by KERIS. KERIS has been publishing 1,981,255 bibliographic data, and 8,143 name authority data as shown in Table 1.

NLK has been publishing NLK LOD (https://lod.nl.go.kr) of bibliographic records, Name authority data, and Subject headings in early 2014 with linking external LOD. It has used properties and classes from lots of ontology such as DC, DCTERMS, BIBO, BIBFRAME, FOAF, SKOS, and NLK defined local ontology. Name authority data were converted to LOD using FOAF, and National Library of Subject Headings (NLSH) were transformed to LOD according to SKOS. ISNI, LDS, and VIAF have been consumed for linking vocabularies of external LOD by NLK. NLK has been publishing 19,775,931 bibliographic data, 346,888 authority data, and 542,661 subject headings as shown in Table 1.

KISTI has been publishing KISTI LOD (http://lod.ndsl.kr) of bibliographic records for scientific academic information in late 2013. It has used properties and classes from PRISM, DC, Bibtex, FOAF and KISTI devised local ontology. DBpedia, Open Library, Sudoc, and BibBase are most consumed for linking vocabularies of external LODs by KISTI. KISTI has been publishing 1,794,088 bibliographic data focusing on article and 467,574 agent data as shown in Table 1.

TABLE 1: Comparison of 3 LOD services

	KERIS	NLK	KISTI
URI	http://data.riss.kr/serviceHome.do	http://lod.nl.go.kr	http://lod.ndsl.kr
Domain	General	General	Scientific academic information
Volume of Bibliographic data	1,981,255	19,775,931	1,794,088
Volume of Authority data	8,143	346,888	467,574
Volume of Subject	-	542,661	-
Ontology for Bibliographic Data	DC, BIBO MODS	DC, DCTERMS, BIBO BIBFRAME	DC PRISM, Bibtex
Ontology for Agent	FOAF	FOAF	FOAF
Ontology for Subject	SKOS	SKOS	
Local ontology	Keris properties	nlon properties	ndsl properties and classes
Interlinking External LOD	OCLC WorldCat, LCSH, BNB, GeoNames, DBpedia, Flickr	VIAF, LC LDS, ISNI	DBpedia, Open Library, Sudoc, BibBase
Starting Year	late 2013	early 2014	late 2013
This table was based on NIA(2014)			

Ontology

When analyzing ontology used in three LOD, each organization used ontology differently. FOAF for agent (person and organization), and SKOS for subject were used respectively. However, ontology for bibliographic data was various such as DC, BIBO, MODS, DCTERMS, BIBFRAME, PRISM, and Bibtex. KERIS used MODS as well as DC and BIBO, NLK selected BIBFRAME with DC, DCTERMS and BIBO, and KISTI used DC, PRISM, and Bibtex. In particular, NLK adopted BIBFRAME ontology in need of specific properties in transforming bibliographic data to LOD.

In addition to above universal ontology, three LOD developed their own ontology for specific properties and classes because of lacks of classes and properties of standard ontology. KERIS designed its properties for holding information such as keris:institution, keris:library, keris:university, and keris:author as shown in Figure 1. NLK has its properties for local data such as nlon:audienceNote, nlon:supplementNote, nlon:localHolding, nlon:awardsNote and so on as shown in Figure 2. KISTI invented its classes such as ndsl:Article and ndsl:Journal, and its properties such as ndsl:keyword, ndsl:conferenceVenue, and ndsl:yearOfAffiliation as shown in Figure 3.

Three LOD have no choice but to develop their own ontology to transform and publish their legacy data to LOD. However, these local properties and classes would lead to some problems in LOD sharing.

Suggestion for LOD in Korea

These local properties and classes were different with inconsistency that would bring out conflicts in data sharing. In the requirement for metadata, interoperability is very important. Locally developed properties and classes would make data sharing to be difficult because of imperfect crosswalking and mapping.

Among ontological modeling, BIBFRAME is more applicable for library because BIBFRAME reflected FRBR model and accommodated MARC field and subfield to replace MARC. Therefore, for LOD interoperability and crosswalking, this study suggested transforming locally devised ontology of three LOD to BIBFRAME which has been developed as library specific ontology in future.

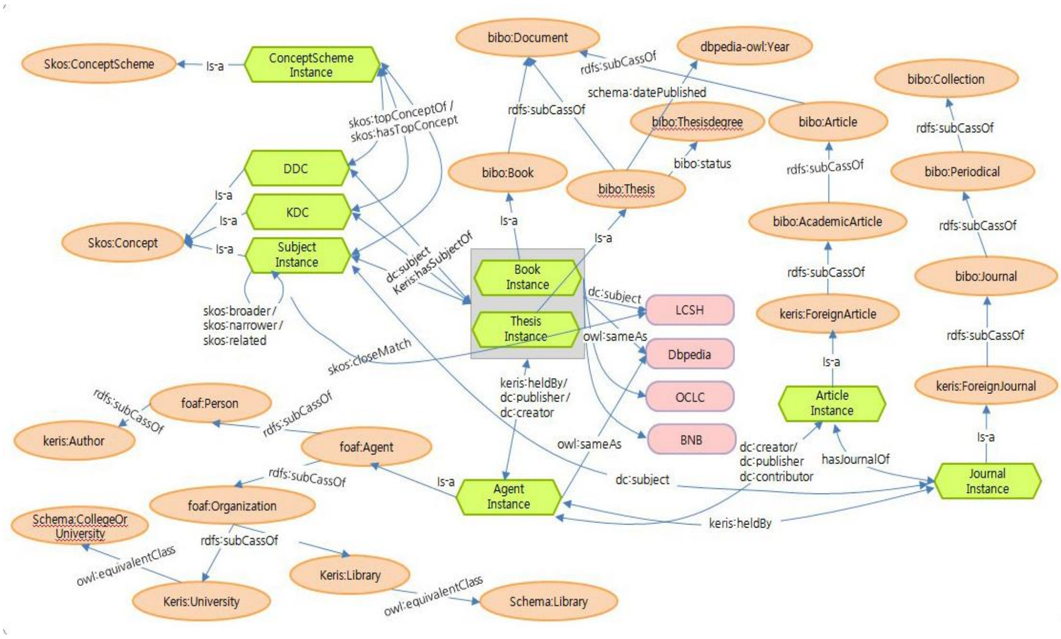


FIG. 1. KERIS ontology
Source: KERIS Home Page

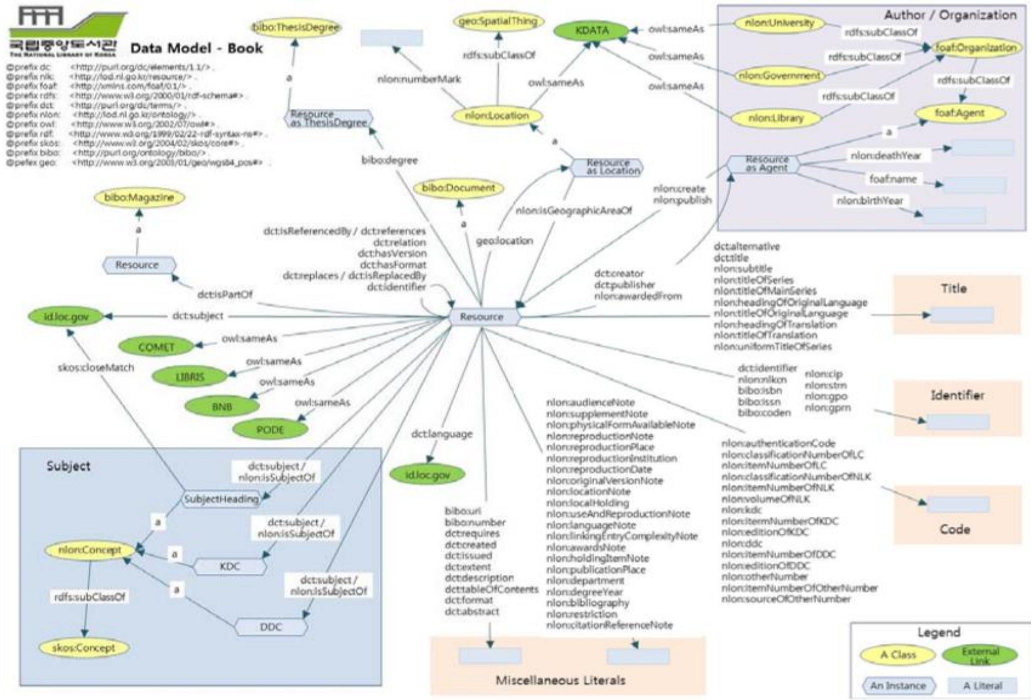


FIG. 2. NLK ontology
Source: NLK Home Page

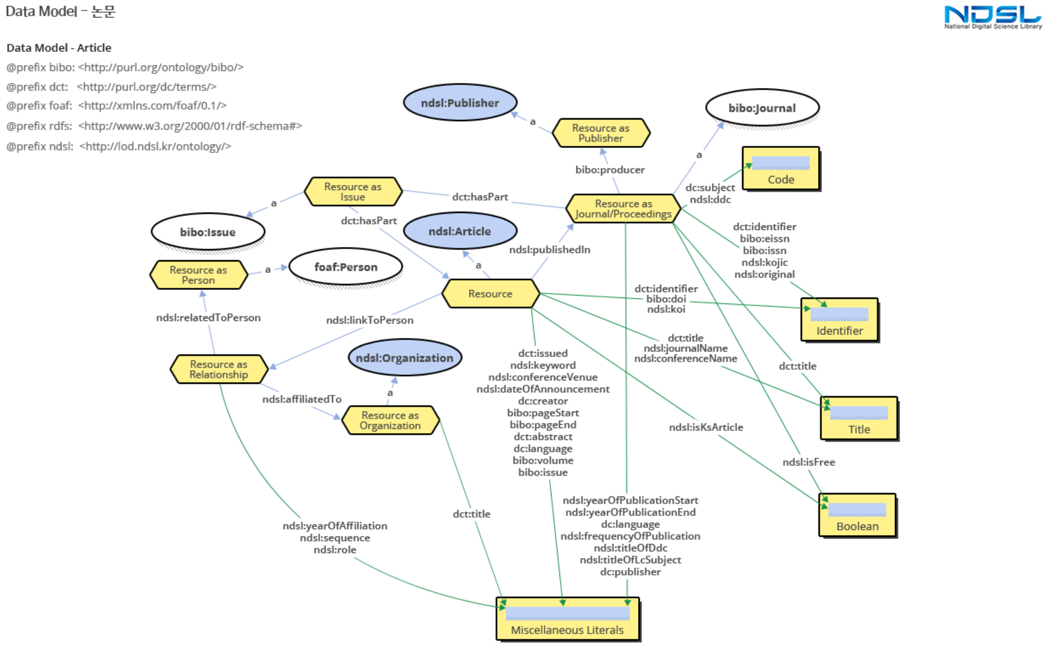


FIG. 3. NDSL ontology (Article resource)
Source: KISTI Home Page

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A8018778)

References

NIA. (2014). Linked Open Data : Korea case. Seoul: NIA.

KERIS Home Page. Retrieved, January 15, 2018, from http://data.riss.kr/LODintro.do.

KISTI Home Page. Retrieved, January 15, 2018, from http://lod.ndsl.kr/home/intro/ontology.jsp.

NLK Home Page. Retrieved, January 15, 2018, from http://lod.nl.go.kr.

Author Identifier Analysis: Name Authority Control in Two Institutional Repositories

Poster

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

Naomi Eichenlaub
Ryerson University, Canada
neichenl@ryerson.ca

Keywords: persistent identifiers; researcher identifiers; public data; name disambiguation; metadata assessment; LCNAF; Scopus Author Identifier; ORCID; VIAF; ISNI.

Abstract

The aim of this poster is to analyze name authority control in two Institutional Repositories (IRs) to determine the extent to which faculty researchers are represented in researcher identifier databases. A purposive sample of 50 faculty authors from Florida Southern College (FSC) and Ryerson University (RU) were compared against five different authority databases: Library of Congress Name Authority File (LCNAF), Scopus, Open Researcher and Contributor ID (ORCID), Virtual International Authority File (VIAF), and International Standard Name Identifier (ISNI). We first analyzed the results locally, then compared them between the two institutions. Looking at both institutions together, the representation was closest in two and three databases (24% and 22% at FSC and 18% and 20% at RU). This has implications for enhancing local authority data by linking to external identifier authority data to augment institutional repository metadata.

Background

Florida Southern College (FSC) is the oldest private college in the state of Florida with 2,500 FTE and 139 full-time faculty members. The college offers undergraduate, graduate, and post-graduate programs in various disciplines. Ryerson University (RU) is located in the heart of downtown Toronto, Canada. With an FTE of 35,000 across more than 100 undergraduate and graduate programs, Ryerson has close to 900 full-time faculty members including 20 Canada Research Chairs. Both institutions have IRs situated within their libraries that contain faculty research.

Metadata quality in IRs has typically been a challenge owing to a number of circumstances. Designed initially for self-submission of faculty research, metadata workflows in IRs are often lacking the functionality to ensure best practices, particularly in the area of authority control (Salo, D., 2009). Moreover, metadata arrives in repositories from a variety of sources including batch ingests, harvesting, and deposits by staff, students and researchers, making it difficult to enforce consistency (Chapman, J.W., Reynolds, D., & Shreeves, S. A., 2009).

Authority control is the process of identifying headings as access points and ensuring each access point is unique by disambiguating variant headings. Traditionally, authority files were created and maintained in-house in libraries and contained unique character strings to identify authorized subject and name headings. In an online environment, the need for authority control has extended beyond the library catalog to IRs and journal article databases. Name authority control in IRs continues to be difficult to manage, especially in larger repositories, and can impact discovery and retrieval as well as attribution. Furthermore, once variations in name headings are introduced to the repository, they most often must be corrected manually, a very time-consuming process (Salo, 2009).

Introduction

In this digital age, computers require more direction than humans in terms of name disambiguation (Van der Graaf, M. & Waaijers, L., 2014). Therefore authority control, and name authority control in particular, becomes even more critical. As such, a number of web-based author identifier initiatives have emerged such as Scopus Author Identifier, LCNAF, VIAF, ISNI, and ORCID. The online environment of IRs offers an opportunity to harness persistent identifier and linked data initiatives and evolve beyond traditional bibliographic data silos to embrace open author identifiers (Chapman, J. W., Reynolds, D., & Shreeves, S.A., 2009), especially as academic institutions place increasing importance on tracking research outputs. A 2013 Program for Cooperative Cataloguing report acknowledged the need to explore "the role of name authorities and identity data generally in a post-MARC, linked data environment" (PCC, p. 4)

To this end, this study looks at name authority control in two separate IRs in the broader context of research identifiers to determine the extent to which faculty are represented in author identifier databases. It builds on previous studies looking at researcher representation in name authority databases (Sandberg, J. & Jin, Q., 2016; Waugh, L., Tarver, H., Phillips, M.E., 2014) in an effort to understand the scope of name authority representation beyond what is currently included in author name metadata in the IR software platforms at Florida Southern College Roux Library and Ryerson University Library and Archives. As an outcome of this study, the authors hope to explore external options available to enhance name authority in their IRs.

Methodology

The authors selected a purposive sample of faculty researchers in their IRs with the goal of surveying the representation of identifiers available for each researcher. Fifty researchers with content from each IR were selected. The careful selection focused on the researchers with enough content to warrant having identifiers in external authority databases and requiring name authority control since they have multiple entries in the IR. The authors then compared this sample of faculty researchers from each IR across five databases: Library of Congress Name Authority File, Scopus, ORCID, VIAF, and ISNI.

For each institution, we created a spreadsheet containing an entry for each author and separate tabs for LCNAF, Scopus Author Identifier, ORCID, VIAF, and ISNI databases. We then manually searched for each name in each database and added the identifiers to the spreadsheet. We did not count ORCID records that were not public, because it was not possible to disambiguate or confirm researcher identify when no public information was available. Once all the data was gathered, we analyzed the coverage of authors represented across databases for both institutions. To calculate the coverage we used the following formula:

$$f = \frac{\sum r}{y}$$

FIG. 1. Formula used to calculate the authors' representation in five databases.

In this formula, r equals the numbers of author representations, and y equals sample size (50).

Results

Reviewing the FSC results, we found that all but 3 authors were represented in a database, with an overall of 86% represented in Scopus, 36% in VIAF, 30% in LCNAF, 30% in ISNI, and 12% in ORCID. The RU results indicate the researcher sample had the strongest author representation in Scopus with 96%, followed by 76% in ISNI, 62% in ORCID, 62% in VIAF and 40% in LCNAF. All but one of the RU researchers were represented in at least one database.

TABLE 1: Percentage of author names represented in each database

	LCNAF	Scopus	ORCID	VIAF	ISNI
Florida Southern College	30%	86%	12%	36%	30%
Ryerson University	40%	96%	62%	62%	76%

Additionally, analyzing the database representation at FSC, we determined that 6% of the authors were not represented in any database, 36% were represented in one database, 24% represented in two databases, 22% in three databases, 12% in four databases, and there was no author representation in all five databases. At RU, the results indicate that 2% of authors were not found in any of the databases, 6% were represented in one database, 18% in two databases, 20% in three databases, 36% in four databases and 18% in five databases. A complete breakdown of authors’ representation is found in table 2.

TABLE 2: Representation of authors in five identifier databases for each institution

	0 DB		**1 DB**		**2 DB**		**3 DB**		**4 DB**		**5 DB**	
	FSC	**RU**	**FSC**	**RU**	**FSC**	**RU**	**FSC**	**RU**	**FSC**	**RU**	**FSC**	**RU**
No database	3	1										
LCNAF			0	0								
Scopus			18	3								
ORCID			0	0								
VIAF			0	0								
ISNI			0	0								
LCNAF and Scopus					0	0						
LCNAF and ORCID					1	0						
LCNAF and VIAF					0	0						
LCNAF and ISNI					0	0						
Scopus and ORCID					4	5						
Scopus and VIAF					3	1						
Scopus and ISNI					5	3						
ORCID and VIAF					0	0						
ORCID and ISNI					0	0						
VIAF and ISNI					0	0						
LCNAF, Scopus, ORCID							0	0				
LCNAF, Scopus, VIAF							6	0				
LCNAF, Scopus, ISNI							0	0				
LCNAF, ORCID, VIAF							0	0				
LCNAF, ORCID, ISNI							0	0				
Scopus, ORCID, VIAF							0	1				
Scopus, ORCID, ISNI							1	7				
Scopus, VIAF, ISNI							0	2				
LCNAF, VIAF, ISNI							4	0				
ORCID, VIAF, ISNI							0	0				
LCNAF, Scopus, ORCID, VIAF									1	1		
LCNAF, Scopus, ORCID, ISNI									0	0		
LCNAF, Scopus, VIAF, ISNI									5	9		
LCNAF, ORCID, VIAF, ISNI									0	1		
Scopus, ORCID, VIAF, ISNI									0	7		
LCNAF, Scopus, ORCID, VIAF, ISNI											0	9
TOTAL %	6%	2%	36%	6%	24%	18%	22%	20%	12%	36%	0%	18%

The overall state of representation of author identifiers in this study shows that the highest representation of authors at FSC was found in one, two and three databases. At RU, the highest representations of authors was found in four databases, three databases, and two and five databases (tied), respectively. Looking at both institutions together, the representation was closest in two and three databases (24% and 22% at FSC and 18% and 20% at RU). Since the sample selected was purposive, a completely random sample of researchers may have yielded different results for each institution. Moreover, the findings show that while the majority of database permutation results are comparable between the two institutions, the difference in results from one database (Scopus), four databases (Scopus, ORCID, VIAF, and ISNI), and five databases (LCNAF, Scopus, ORCID, VIAF, and ISNI) is considerable.

As Sandberg & Jin justified, different disciplines have different representation results (Sandberg, J. & Jin, Q., 2016). Science faculty may score lower on book-centric databases such as LCNAF, and higher in Scopus with content driven from serial publications and conference series.

Future Work

Authority control in IRs must become more of a focus, especially as institutions place a higher priority on tracking researcher outputs. To reduce data silos, name authority control in IRs is an opportunity to harness existing external author identifiers such as ORCID, Scopus Author Identifier, ISNI, VIAF, and LCNAF but we need software and platforms that help us take advantage of these, for example through bi-directional updates. An example of this could be ORCID integration with IRs, which serves both to increase the number of researchers with ORCID iDs as well as to match ORCID iDs with institutional affiliation.

In a global research environment, IR managers must prioritize name authority work and advocate for increased system functionality to help manage it, not only IR functionality such as batch edits, global updates and auto-complete (Salo, D., 2009), which do little to enforce authority control, but also to take advantage of linked data resources to maximize interoperability and showcase researcher output.

Conclusions

Authority control in IRs can no longer be limited to manual, in-house cleanup of name headings, but most also leverage the authority data that currently exists in external author identifier sources such as ORCID, Scopus, ISNI, VIAF, and LCNAF, to collectively work together to confirm name identity. The authors determined via a purposive sample that in analyzing the author identifier representation of 100 researchers across two institutions, the results are closest for the two institutions in 2 and 3 databases: 24% and 22% for FSC and 18% and 20% for RU. As a result of these findings, the authors will look at ways of leveraging external name authorities across various external platforms to enhance name authority control in their IRs.

References

Chapman, John W., David Reynolds, and Sarah A. Shreeves. (2009). Repository metadata: Approaches and challenges. Cataloging & Classification Quarterly, 47(3-4), 309-325. 10.1080/01639370902735020

Report for PCC Task Group on the Creation and Function of Name Authorities in a Non-MARC Environment. (2013). http://www.loc.gov/aba/pcc/rda/RDA%20Task%20groups%20and%20charges/ReportPCCTGonNameAuthInA_NonMARC_Environ_FinalReport.pdf

Salo, Dorothea. (2009). Name authority control in institutional repositories. Cataloging & Classification Quarterly, 47(3-4), 249-261. 10.1080/01639370902737232

Sandberg, Jane and Qiang Jin. (2016). How should catalogers provide authority control for journal article authors? Name identifiers in the linked data world. Cataloging & Classification Quarterly, 54(8), 537-552. 10.1080/01639374.2016.1238429

Van der Graaf, Mauritz and Leo Waaijers. (2014). Authority files: breaking out of the library silo to become signposts for research information. http://repository.jisc.ac.uk/6224/1/Authority_files_-_Breaking_out_of_the_library_silo.pdf

Waugh, Laura, Hannah Tarver, and Mark Edward Phillips. (2014). Introducing name authority into an ETD collection. Library management 35 (4 / 5). 10.1108/LM-08-2013-0074

Visualizing Library Metadata for Discovery
Poster

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

Stephanie R. Baker
University of Illinois at Urbana-Champaign, USA
srbaker@illinois.edu

Peiyuan Zhao
University of Illinois at Urbana-Champaign, USA
pzhao12@illinois.edu

Jiawei Li
University of Illinois at Urbana-Champaign, USA
jiaweil3@illinois.edu

Keywords: metadata; discovery; visualization; catalog records; data structure; MARC 21

Abstract

The benefits of visualization have been discussed widely and it is already implemented into library services. However, use cases for visualization have been mostly focused on collection analysis to improve collection development policies and budget management, not for discovery services that take full advantage of the rich information contained in library catalog records.

One of the challenges of working with library catalog records for visualization is the sheer volume of elements (such as control field, data field, subfield, and indicators) and information included in the MAchine-Readable Cataloging (MARC) format records. As is well-known, there are more than 1,900 fields in the MARC 21, which is just too many to use for effective visualizations (Moen and Benardino, 2003). In addition, some fields are used for recording the same information, for example, the control field 008 positions 7 to 14 and the subfield $c of the data field 264 are used for the production related date information. Instead of showing a clear relationship between resources, the large number of elements and duplicated information included in the catalog record may muddle those relationships in any visualization. The question then is which information added in which fields of the MARC 21 format catalog records should be considered essential information to be included in library catalog data visualizations for discovery.

According to Mischo, Schlembach, and Norman's research (2009) on users' search query terms analysis, users tend to use more than three words as search terms (i.e., known item search) rather than simple keyword searches. Many users also use full citations as search terms, thus showing that library users very often already know what they want when they come to the library gateway. Consequently, for the purpose of supporting Functional Requirement for Bibliographic Record (FRBR) User Tasks (IFLA, 2017), such as finding, identifying, selecting, and acquiring (along with browsing), library discovery service systems *do not* need to index all of the elements included in MARC 21 format catalog records. What is needed instead is only the key information that affects the discovery services, such as access point and authorized access point that connect FRBR Group 1 entities (e.g., work, expression, manifestation, and items) defined by the Resource Description and Access standards (Library of Congress, 2017).

TABLE 1. Data used for the prototype discovery service that employed visualization tool.

Data	MARC data fields	Data	MARC data fields
Name (Agent)	100, 110, 111, 700, 710, 711	Subject	050, 082
Title (Work)	130, 245, 246	Date	260 $c or 264 $c
Bibliographic record identifier	001 (Local bibliographic record ID)	Holdings record identifier	004 (Local holdings record ID)

Since visualizations work best for showing relationships between resources, the researchers at the University of Illinois at Urbana-Champaign Library developed entity relationships between 'work (title),' 'name' and 'subject.' Those relationships are displayed through visualizations that provide opportunities for users to understand, identify, and find related and similar resources in a more effective and organized manner. The information used for these relationships was extracted from a sample of 300,000 randomly selected library catalog records (from 7.4 million total catalog records) as shown in table 1 above. A prototype discovery service that employs the visualization tool D3.js (https://d3js.org/) was created.

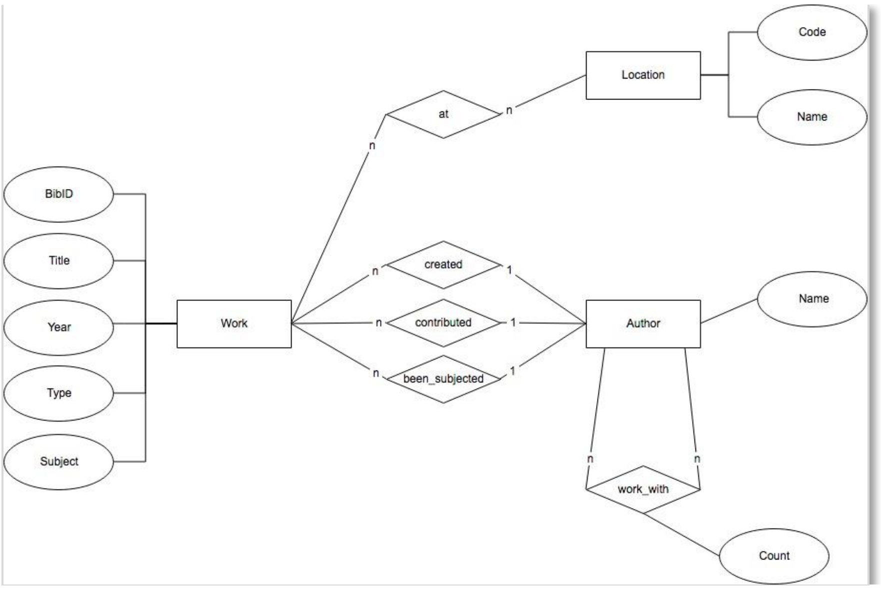
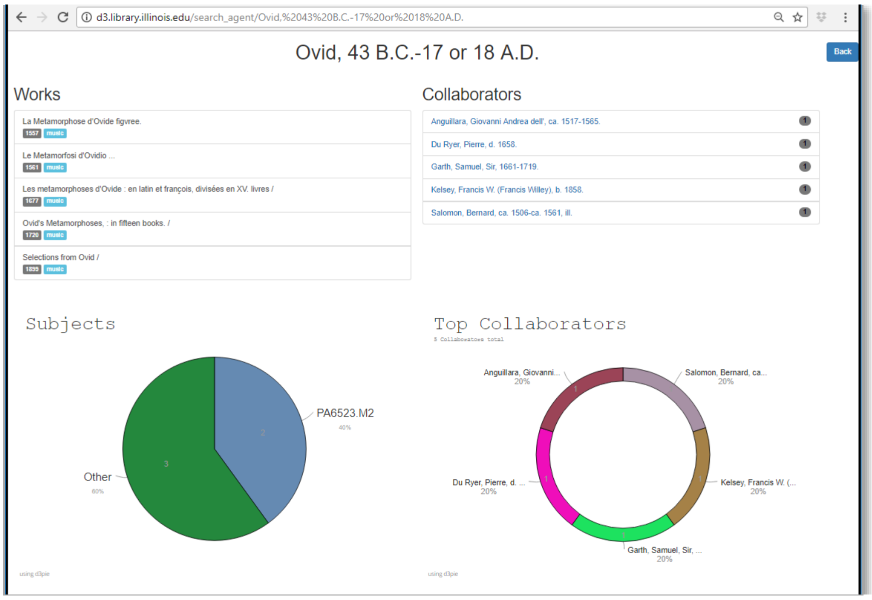


FIG. 1. Entity relationships diagram for the work search.

The new prototype discovery service supports only two simple search options, work and name, with the assumption that users will perform a 'known item search,' as mentioned earlier. The search results page displays related resources by visualizing relationships between entities as shown in figure 1. For example, if a user starts the search with a title, then the result page allows the user to browse related resources (works) by the same author, on the same topic, with the same publication date, or having the same holdings library(ies). If a user starts the search with a name, the search results page allows the user to browse by relationships associated with the name, such as works created by the name, close collaborators, and subject areas of all works associated with the name. Figure 2 shows an example of the resulting page from a name search.

This is different from the current faceted browsing services provided in many existing library discovery services. Instead of displaying numbers of items with the same information as a list, the prototype discovery service combines the results together and displays them as a visualization. When the user selects the item (work), then the prototype discovery service displays all indexed information including holdings libraries. On the same page, the prototype provides a link to the full catalog record page in case the user wants to see all of the information included in the MARC 21 format catalog record that may help users to identify and select the resource.

This experimentation confirmed that visualizing library catalog data is not that easy even with small set of data from a sample records. The challenges include inconsistent terms used in the records, data quality, and granularity of data in certain MARC21 data fields. Although we decided to select the data associated with the access points for this experimentation, notions of what constitutes key information for discovery services is not clear yet. However, the prototype showed the benefits of using a selective set of data critical to discovery and visualization, as opposed to using all of the information included in an entire catalog record.



d3.library.illinois.edu/search_agent/Ovid,%2043%20B.C.-17%20or%2018%20A.D.

Ovid, 43 B.C.-17 or 18 A.D.

Back

Works

La Metamorphose d'Ovide figvree.
1557 music

Le Metamorfosi d'Ovidio ...
1561 music

Les metamorphoses d'Ovide : en latin et françois, divisées en XV. livres /
1677 music

Ovid's Metamorphoses, : in fifteen books. /
1720 music

Selections from Ovid /
1899 music

Collaborators

Anguillara, Giovanni Andrea dell', ca. 1517-1565.	1
Du Ryer, Pierre, d. 1658.	1
Garth, Samuel, Sir, 1661-1719.	1
Kelsey, Francis W. (Francis Willey), b. 1858.	1
Salomon, Bernard, ca. 1506-ca. 1561, ill.	1

FIG. 2. A search result page that shows works, subject and collaborators related with the name.

For the next step, we will work with a complete set of library catalog records to test the full functionality of the service and the impact of the visualization. It will also include using URIs of linked data sources for entities that would significantly improve the library's visualization-based discovery service. With the maturing of the BIBFRAME ontology, our project will use the set of BIBFRAME vocabularies describing entities, such as work, instance, item, agents, subjects, and events (Library of Congress, 2016) for visualization by adapting workflows established from this experimentation. We also hope that a proper user testing should be conducted to identify an ideal set of bibliographic data used for the discovery services.

References

International Federation of Library Associations and Institutions (IFLA). (2017). Functional Requirements for Bibliographic Records (FRBR). Retrieved, May 5, 2018, from https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8915.

Library of Congress. (2016). Overview of the BIBFRAME 2.0 Model. Retrieved, July 17, 2018, from https://www.loc.gov/bibframe/docs/bibframe2-model.html.

Library of Congress. (2017). Resource Description and Access (RDA). Retrieved, May 5, 2018, from https://www.loc.gov/aba/rda/.

Mischo, William. H., Mary Schlembach, Michael Norman. (2009). Modeling search assistance mechanisms within web-scale discovery systems. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pp. 431.

Moen, William E., and Penelope Benardino. (2003). Assessing Metadata Utilization: An Analysis of MARC Content Designation Use. Proceedings of the 2003 Dublin Core International Conference. Retrieved, May 5, 2018, from http://dcpapers.dublincore.org/pubs/article/view/745/741.

Building a Framework to Encourage the use of Metadata in Modern Web-Design
Poster

Jackson Morgan
Georgia Institute of Technology, United States of America
jmorgan45@gatech.edu

Keywords: javascript, event-driven, framework, web development, rdf, sparql

Abstract

When Tim Berners-Lee published the roadmap for the semantic web in 1998, it was a promising glimpse into what could be accomplished with a standardized metadata system, but nearly 20 years later, adoption of the semantic web has been less than stellar. In those years, web technology has changed drastically, and techniques for implementing semantic web compliant sites have become relatively inaccessible. This poster outlines a JavaScript framework called Beltline.js which seeks to encourage the use of metadata by making it easy to integrate into modern web best-practices.

Introduction

As interactive websites have become more ubiquitous, JavaScript has increasingly been a must-know language for web developers. It has become a household name for developing applications becoming by far the most used language on github making up 15% of the site at over 300,000 repositories (Zapponi, 2014).

Because of the popularity of JavaScript, libraries targeted at JavaScript developers have been able to introduce novel concepts to a wide audience. One such library that would be of interest to the semantic web community is Facebook's GraphQL (GraphQL, 2018). GraphQL provides a query language that allows a client to define the structure of information it wants to receive before querying the server. In a way, it's reminiscent of a few features in the Resource Description Framework (RDF) and its query language SPARQL. However, GraphQL is not designed to further the availability of metadata. Unlike RDF which is designed to link many domains together with triples, GraphQL is built around the traditional idea of isolated servers providing domain specific data. Nonetheless, because GraphQL is a JavaScript framework, a new generation of developers was introduced to the concept of graph-based data.

Beltline.js is a JavaScript library that aims to capitalize on the viral nature of JavaScript libraries while increasing the adoption of RDF. Named after the Beltline in Atlanta Georgia, an ambitious path and transit corridor that will connect a plethora of neighborhoods in a shared travel experience, Beltline aims to connect the various aspects of a JavaScript web application through a shared data interface.

Architectural Inspiration

Event-driven architectures have become increasingly popular among web developers. By utilizing multi-directional communication, oftentimes implemented with WebSockets, a developer is able to build an experience that keeps a user interface up-to-date with the application as a whole. Almost all interfaces that push updates to users without requiring a user to reload, navigate to a new page, or interact with the UI are supported event-driven architectures. (Michelson, 2011)

Beltline is heavily influenced by the Distributed Data Protocol (DDP), a standard for event driven architectures most known for its use in the popular JavaScript framework, Meteor.js. (DDP, 2016) While originally designed to integrate with the database MongoDB, with a few modifications, DDP can work with triplestore databases. Beltline's implementation is designed to accomplish 3 architectural goals in order to achieve developer ease:

Event-Driven: It should not rely on a request-reply architecture

Controllable: A developer should be able to easily control how much data a client is given

Integrable: It should be easy for a developer to integrate Beltline into their current tech-stack.

Implementation

Beltline's *event-driven* architecture is enabled by WebSockets. When a web browser connects to a Beltline-enabled web site, it makes a WebSocket (WebSockets, 2018) connection with the server using Socket.io (Socket.io, 2018). This enables not only a client's ability to push data to the server but the server's ability to push data directly to its clients unprompted.

Beltline's method of syncing data between the client and the server takes heavy inspirations from Meteor's Mini-Mongo (MiniMongo, 2018), but using the JavaScript triplestore, rdfstore-js (rdfstore-js, 2018), a JavaScript library that has the same interface and functionality as a normal triplestore database. As it is completely built in JavaScript, our JavaScript triplestore can be instantiated inside a web page upon load. This allows a user to query a database locally as if the database were on the same machine. Beltline keeps each instance of the JavaScript triplestore across all connected web pages in sync with the main triplestore database on a developer's server. When a request is made to update the database using Beltline's call method, Beltline follows the optimistic UI design pattern (Stubailo, 2015). The request first updates the JavaScript triplestore as if nothing went wrong. It then makes the request to the server, and when the request is properly received, it updates all other clients with the new information.

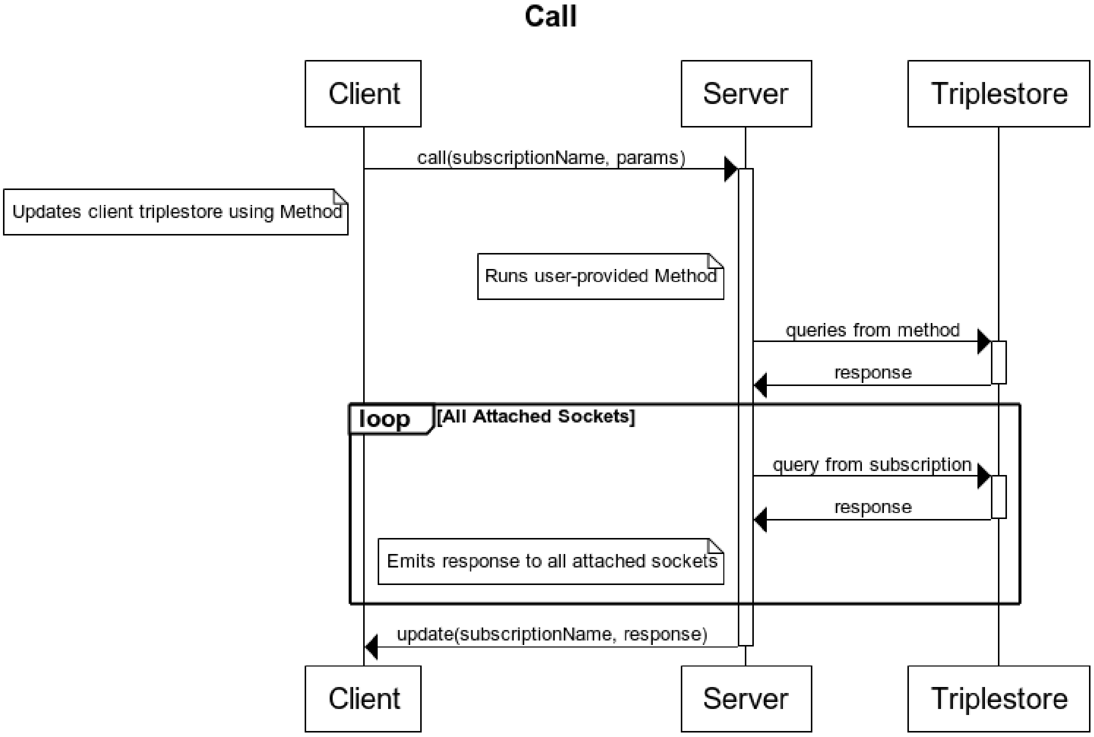

Call
Client
Server
Triplestore
call(subscriptionName, params)
Updates client triplestore using Method
Runs user-provided Method
queries from method
response
loop
[All Attached Sockets]
query from subscription
response
Emits response to all attached sockets
update(subscriptionName, response)
Client
Server
Triplestore

FIG. 1. Sequence for making a database update.

One major downfall of replicating a database across many clients is that it is not *controllable.* There are very few applications that would be designed to load the entirety of a database into a web page. Fortunately, DDP offers a solution to this conundrum in the form of the publish and subscribe methods. Beltline follows suit, employing the same functionality and method names.

Beltline's publish method lives on the server and accepts an id to denote what is being published by taking a user-defined function as parameters. On the client, the subscribe method can be called by passing in an id corresponding to one of the publish functions. At this point, the function will run a SPARQL query on the server, extracting the desired data. That data is then sent as triples to the client to be stored in the client's JavaScript triplestore. It should be noted that these methods only work with CONSTRUCT SPARQL queries as the result must be a graph to be shared between the server and client.

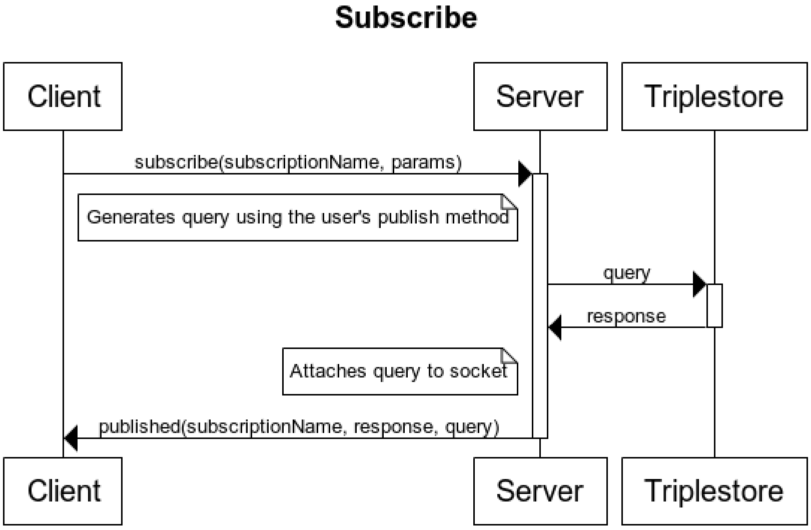

Subscribe
Client
Server
Triplestore
subscribe(subscriptionName, params)
Generates query using the user's publish method
query
response
Attaches query to socket
published(subscriptionName, response, query)
Client
Server
Triplestore

FIG. 2. Sequence for subscribing and publishing.

Finally, in order to increase adoption, Beltline must be *integrable* in frameworks with which developers are familiar. To integrate Beltline, a developer simply needs to import Beltline's server middleware, configure it, and attach it to a route using Express (a popular JavaScript server framework) or some other server framework. Configuration options include the IP address of the main database and optional database plugins. Plugins serve the purpose of extending the usability of Beltline to developers who do not use a SPARQL compatible database. For example, one plugin could transform Beltline's SPARQL queries to SQL queries to hook into a developer's SQL database (Prud'hommeaux, 2009).

Because there is a wide array of client-side JavaScript frameworks, Beltline's frontend libraries must be framework agnostic. As a result, Beltline relies on callback functions that can be passed into the subscribe or call methods at any point in the code. When an event happens, these functions will be triggered, which could lead to an update anywhere else in the client-side codebase.

Beltline not only provides a convenient solution for developers hoping to build event-driven web applications, it also encourages them to open a SPARQL endpoint for their data. Often, making a site semantic web compliant as an extraneous task for the developer, and her effort could be better spent developing new features. Beltline makes feature development and the exposure of semantic data one in the same by making metadata core to a JavaScript framework.

A demo of Beltline can be viewed at https://github.com/jaxoncreed/beltline-example and the full implementation can be installed at https://www.npmjs.com/package/beltline.

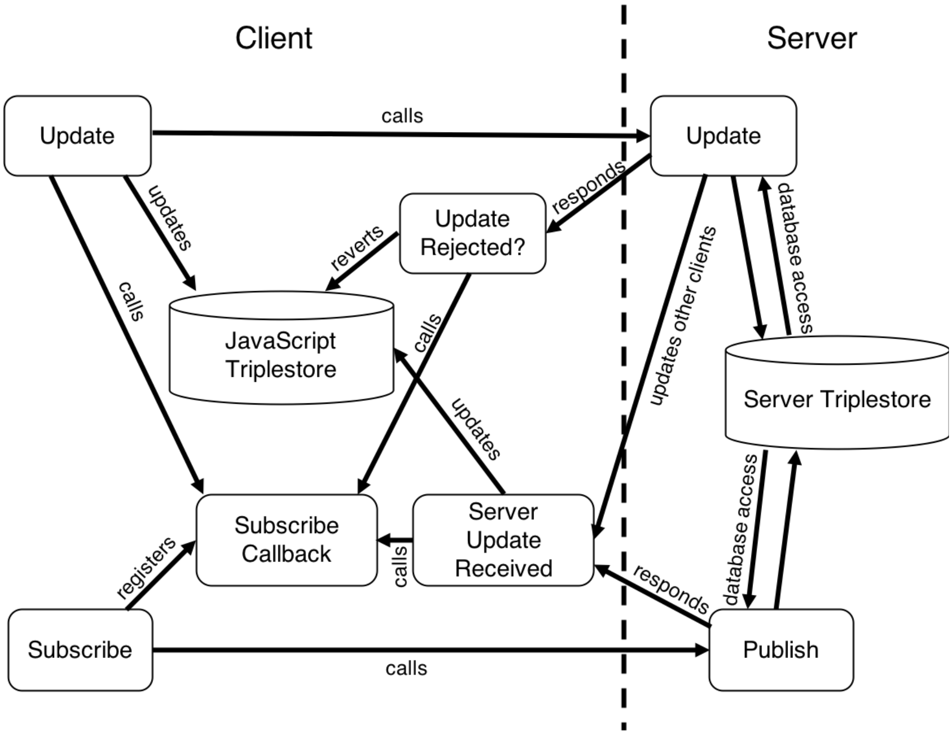


FIG. 3. Beltline Data-Flow.

Conclusion

Beltline not only provides a convenient solution for developers hoping to build event-driven web applications, it also encourages them to open a SPARQL endpoint for their data. Often, making a site semantic web compliant as an extraneous task for the developer, and her effort could be better spent developing new features. Beltline makes feature development and the exposure of semantic data one in the same by making metadata core to a JavaScript framework.

References

DDP. Retrieved, August 16, 2018, from https://github.com/meteor/meteor/blob/devel/packages/ddp/DDP.md.

GraphQL. Retrieved, May 6, 2018, from https://graphql.org.

Meteor. Retrieved, May 6, 2018, from https://www.meteor.com.

Michelson, Brenda M. (2011, February). Event-Driven Architecture Overview. Elemental Links. Retrieved, May 6, 2018, from http://elementallinks.com/el-reports/EventDrivenArchitectureOverview_ElementalLinks_Feb2011.pdf.

MiniMongo. Retrieved, May 6, 2018, from https://github.com/mWater/minimongo.

Prud'hommeaux, Eric, and Alexandre Bertails. (2009). A mapping of sparql onto conventional sql. World Wide Web Consortium (W3C). Retrieved, May 6, 2018, from https://www.w3.org/2008/07/MappingRules/StemMapping.

rdfstore-js. Retrieved, May 6, 2018, from https://github.com/antoniogarrote/rdfstore-js.

Socket.io. Retrieved, May 6, 2018, from https://socket.io.

Stubailo, Sashko. (2015, May). Optimistic UI with Meteor. Retrieved, May 6, 2018, from https://blog.meteor.com/optimistic-ui-with-meteor-67b5a78c3fcf.

WebSockets. (2018, March). Mozilla Developer Network. Retrieved, May 6, 2018, from https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API.

Zapponi, Carlo. (2014). GitHut – Programming Languages and Github. Retrieved, May 6, 2018, from http://githut.info

Analysis of user-supplied metadata in a health sciences institutional repository

Poster

Joelen Pastva
Galter Health Sciences Library &
Learning Center, Northwestern
University, USA
joelen.pastva@northwestern.edu

Keywords: metadata; institutional repository; interface design; health sciences

Abstract

Launched in October, 2015 by the Galter Health Sciences Library, the DigitalHub repository is designed to capture and preserve the scholarly outputs of Northwestern Medicine. A major motivation to deposit in the repository is the possibility of improved citations and discovery of resources, especially for non-traditional materials such as poster presentations and teaching resources that are typically never made publicly accessible.

One of the largest barriers hampering discovery is a lack of descriptive metadata. DigitalHub was designed for ease of use for the depositor, requiring very minimal metadata in order to successfully deposit a resource. However, many optional descriptive metadata fields are also made available, some using auto-complete suggestions from controlled vocabularies wherever possible to encourage the consistent and detailed entry of descriptive information. Although the library can deposit materials on behalf of researchers, the repository is largely intended for the self-deposit of items by researchers. In an effort to improve the discoverability of resources deposited in DigitalHub, the Collection Management and Metadata Services department at Galter Library provides metadata enhancement services for all publicly accessible items. However, the library was curious to evaluate how users were approaching available metadata fields and accompanying instructions prior to the performance of enhancement operations.

In order to evaluate user-supplied metadata, an export was made of all of the metadata in DigitalHub for a 2.5 year period. Records previously enhanced by librarians, or records initially deposited by library staff were excluded from primary consideration. The metadata was then evaluated for completeness, choice of dropdown terms for resource type, inclusion of collaborators, use of controlled vocabulary fields, and any areas that indicated a clear misunderstanding of the intended use of the metadata field. This poster presents the preliminary findings of this analysis of user-supplied metadata.

Although all fields were used appropriately by depositors, over half of all optional metadata fields were left blank, with another 25% of optional fields underutilized. It was especially interesting to observe no use of the Contributor field, although depositors did often record multiple authors. 38% of depositors used a filename for a resource title, which is supplied by the repository by default upon deposit. Depositors were comfortable supplying their own keyword tags, but never utilized auto-suggested controlled vocabulary terms such as LCSH or MeSH for indexing. Despite a rich offering of nearly 160 resource types to accommodate different outputs, only 17 unique resource types were selected by depositors over 72 individual deposits.

It is hoped that the findings of this analysis will help guide future system and interface design decisions, cleanup activities, and library instruction activities. The lack of complete metadata supplied by depositors indicates the continued need for library metadata enhancement for improved discovery. There are also opportunities for the system to pre-populate fields that tend to

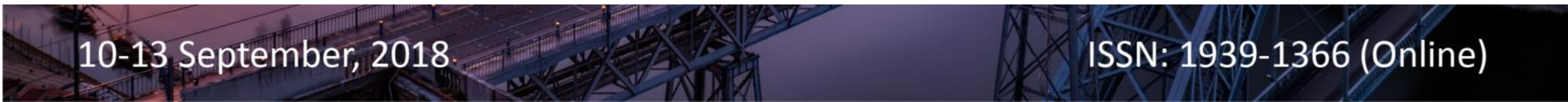
be standardized across all records to improve the richness of resource description upon deposit. Ultimately the goal is to make the interface as usable and effective as possible to encourage depositors to supply an optimal amount of descriptive metadata upfront, and to continue using the repository in the future. These results should be of interest to repository managers that rely on users to supply initial descriptive metadata, especially for health sciences disciplines.

DC-2018
Porto, Portugal

2018 Proceedings of the

International Conference on Dublin Core and Metadata Applications

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



10-13 September, 2018

ISSN: 1939-1366 (Online)