Advancing Metadata Practice: Quality, Openness, Interoperability

2017 Proceedings of the International Conference on Dublin Core and Metadata Applications

26-29 October 2017

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

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-2017
Welcome

Welcome to DCMI 2017, in Crystal City, Virginia!

Having recently celebrated its twentieth anniversary and as the Dublin Core Metadata Initiative (DCMI) enters its third decade, it seems that the importance of metadata research, innovation and practice is undiminished. As an awareness of metadata becomes more mainstream, so the challenges associated with its development and management become more pressing. Looking at this year's conference programme, with peers reporting on a full range of cutting-edge innovation, ongoing development and good practice, I think you will agree that the DCMI community is rising to that challenge!

Your participation in the community's annual meeting and conference gives you a chance to rub shoulders with like-minded people, with experts in research, development and practice. DCMI is all about the community, and the contribution of people like you. As you enjoy this year's conference, we hope that you will consider how you can contribute to the ongoing success of DCMI by participating in one of the committees, or by joining as a member (either individually or as an organisation). A thriving DCMI benefits us all.

I hope that you will also take the opportunity to attend the Open Community meeting on Sunday, where you can meet some of the people active in DCMI work, and where you are invited to contribute your ideas for the future of our community

Finally, I hope that you enjoy the conference, that you manage to engage in interesting and fruitful conversation, and that you leave with ideas and inspiration for another year of working with metadata!

Paul Walk, Chair, DCMI Governing Board

Program Committee Chairs' Welcome

In reflecting on the 2017 edition of the Dublin Core conference, we'll start out by invoking what has become a cliché, the Linked Open Data Cloud. But our community benefits from an occasional reflection on what this image reveals. The current version, published in August 2017, shows a dramatic change from the previous version published in 2014. The cloud is much denser, of course, but the pace of change has also accelerated, and now represents domains that were barely visible in earlier years, including life sciences, geography, and many cross-domain topics. In the library community, interest in structured data and semantic web technology is expanding beyond traditional areas of interest such as cultural heritage collections. New mandates are driving increased interest in institutional repositories and other digital collections outside the library domain. And as we reflect on where this expansion is occurring, it is important to understand how and why these domains are expecting to leverage the technology and research that the DCMI community supports.

Following up on DC-2016 in Copenhagen, with its look at the changing role of metadata in the Second Machine Age, DC-2017 concentrates on evolving technologies and practices that are advancing how we create and manage quality, actionable metadata. Interoperability and openness have been guiding principles of the DCMI community for over twenty years, and these principles have evolved through the development of Semantic Web standards and Linked Open Data. A deluge of new data sources is magnifying the perennial challenge of metadata quality, and is now inspiring the development of innovative tools, practices, and solutions. This year's conference shows some of the possibilities when interoperability is supercharged.

With a variety of presentations, workshops, and demonstrations, the proceedings reflect the interdisciplinary nature of metadata work. The program committee co-chairs are proud to continue DCMI's tradition of bringing together a unique blend of metadata practitioners, researchers, standards developers, and visionaries to share and interact and continue the conversation.

We thank all individuals and teams who submitted proposals to DC-2017. We are grateful to the Dublin Core community for continuing to serve as a volunteer program committee, which has been incredibly responsive to our requests to review submissions and shape the conference. And as we prepare to travel to the meeting itself, we are looking forward to the chance to meet all of you.

Carol Jean Godby, OCLC Research

Michael Lauruhn, Elsevier

ORGANIZING COMMITTEE

Conference Co-Chairs

Stuart A. Sutton, University of Washington, United States
Paul Walk, Dublin Core Metadata Initiative (DCMI) & Antleaf, Ltd., United Kingdom

Program Committee Co-Chairs

Carol Jean Godby, OCLC, United States
Michael Lauruhn, Elsevier, United Stat***es***

Program Committee

Leif Andresen, Royal Danish Library, Denmark
Thomas Baker, Dublin Core Metadata Initiative (DCMI), Germany
Ana Alice Baptista, Universidade do Minho, Portugal
Uldis Bojars, National Library of Latvia, Latvia
Michael Robert Bolam, University of Pittsburgh, United States
Dan Brickley, Vrije Universiteit Amsterdam
Joseph A. Busch, Taxonomy Strategies, United States
Barbara Bushman, National Library of Medicine
Michael D. Crandall, University of Washington, United States
Makx Dekkers, Independent Consultant, Spain
Corine Deliot, British Library, United Kingdom
Gordon Dunsire, Independent Consultant, United Kingdom
Jane Greenberg, Drexel University, United States
Corey A. Harper, Elsevier Labs, United States
Diane Ileana Hillmann, Metadata Management Associates LLC, United States
Eero Hyvönen, Aalto University, Finland
Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Netherlands
Masahide Kanzaki, Keio University Xenon Limited Partners, Japan
Wouter Klapwijk, Stellenbosch University, South Africa
Akira Maeda, Ritsumeikan University, Japan
Mariana Curado Malta, CEOS.PP - Polytechnic of Oporto, Portugal
Deborah Maron, UNC Chapel Hill, United States
Filiberto Felipe Martinez-Arellano, National Autonomus University of Mexico, Mexico
Shawne Miksa, University of North Texas, United States
Peter E Murray, Index Data, United States
Jin-Cheon Na, Nanyang Technological University, Singapore
Johan Oomen, Netherlands Institute for Sound and Vision, Netherlands
Oknam Park, Sangmyung University, Republic of Korea, Korea, Republic Of
Cristina Pattuelli, Pratt Institute, United States
Susanna Peruginelli, Susanna Peruginelli Library consultancy, Italy
Jess Peterson, Amazon, United States
Vivien Petras, Humboldt-Universität zu Berlin, Germany
Magnus Pfeffer, Stuttgart Media University, Germany
Sarah Potvin, Texas A&M University Libraries, United States
Jian Qin, Syracuse University, United States
John Roberts, Archives of Ontario, Canada
Stefanie Ruehle, SUB Goettingen, Germany
David Talley, University of Washington & Lyons Consulting Group, United States
Johann Wanja Schaible, GESIS - Leibniz-Institute for the Social Sciences, Germany
Ryan Shaw, University of North Carolina at Chapel Hill, United States
Lars G. Svensson, Deutsche Nationalbibliothek, Germany
Hannah Tarver, University of North Texas Libraries, United States
Joseph T. Tennis, University of Washington, United States
Anna Tordai, Elsevier, Netherlands
Douglas Tudhope, University of Glamorgan, United Kingdom
Paul Walk, Dublin Core Metadata Initiative (DCMI) & Antleaf, Ltd., United Kingdom
Shenghui Wang, OCLC Research, United States
Oksana Zavalina, University of North Texas, United States
Marcia Lei Zeng, Kent State University, United States

TABLE OF CONTENTS

AUTHOR INDEX

Metadata Theory and Practice

'More Than What It Seems': How Critical Theory, Popular Engagement and Apps Like Tinder Can Help Us Reframe Metadata and Its Consequences

Deborah Maron
University of North Carolina, NC, USA
maron@live.unc.edu

Erin Carter
Cisco Systems
erincarter@alumni.unc.edu

Abstract

Metadata is a term no longer only of interest to information professionals; recently, it has also compelled a wider global population. How might the metadata community guide popular understandings around metadata's relationship to privacy, surveillance, and identity building, while also taking cues from the outside to complement current professional practice? Rather than taking at face value the definitions, presentations, skills, practices and situations that we are told constitute the concept of metadata, we can consider alternative and complementary thinking, broadening what we consider to be metadata at all; this process of rethinking is known as *problematization* and has its roots in critical theory. We use problematization, as well as critical theory constructs like Derrida's *différance* and *digital trace,* to examine the popular dating site Tinder, which we consider to be metadata in its own right. In doing so, we make new assumptions about metadata and its implications in digitally-mediated, surveilled identity construction. We hope that our effort—a contribution to Science and Technology Studies (STS) and also to metadata studies—has professional implications, such as providing companion methods for reading metadata-dependent systems as 'material metadata discourse.' We likewise hope to show that popular, wider-world discourse can cast back onto our profession in a meaningful way.

Keywords: metadata; critical theory; Jacques Derrida; différance; Tinder; digital trace; Bruno Latour; social media; materiality; philosophy

1. Introduction

If a person has remained engaged with the library and information field since the 1980s then they are likely acquainted with the term 'metadata'—at the very least, in the simplified 'data about data' sense. Metadata is a term most familiar to individuals working in descriptive or Linked Data arenas, due in part to the formation of an individuated 'metadata community' developing standards like the Dublin Core, MODS, METS, and auxiliary standards like the Resource Description Framework (RDF) (Coyle, 2005; Harper, 2010). A professional notion of metadata, relating to standards and their encoding schemes, endures in the individuated 'metadata community,' but has been augmented by recent popular 'outsider' interest in metadata, an interest precipitated by recent events relating to metadata's role in issues of privacy and surveillance. We are now ushered into an age where information professionals and laypeople alike are compelled by 'metadata.' It has become so pervasive, we might be living not just in a world *with* metadata but in what Claudio Celis calls a society *of* metadata (Celis, 2015).

In the metadata community, we have our professional understandings of what metadata is, and this influences how we analyze systems using metadata. Typically this involves assessing something like a digital library for its adherence to rules of a standard, like the Dublin Core, via empirical means. Such approaches or methods tell us implicitly how much the metadata dictums of findability, organization and clear description are valued in a library or archive. Though empirical research methods are rigorous and useful to us, we wonder if 'metadata systems' might

benefit from a complementary, qualitative critical treatment that elucidates the ethical, feminist, or identity-forming consequences of their metadata schema implementations. Although there are limited examples from the 'metadata community' that give such qualitatively critical treatments to metadata systems, there are plenty from the 'wider world' as a result of popular media's metadata maelstrom. What if we take cues from the wider world, letting professional metadata 'learn' from the popular interest? A branch of philosophy called *critical theory*, specifically its notion of *problematization*—or the act of turning a practice or thing into a critical object of study—can provide methods of alternative analysis while we as metadata professionals assimilate ideas and concerns from 'popular' conceptions about metadata.

Though there are many popular examples illustrating metadata's critical potentials, we choose to problematize how metadata constructs the virtual identity and relative value of people in the dating application Tinder, which can play out in inequitable ways. Our twofold contribution is to 1) contribute to Science and Technology Studies as well as Metadata Studies 'from the inside,' by asserting critical theory as a useful apparatus for examining metadata and 2) offer complementary methods—specifically, the apparatus *différance and digital trace*—for problematizing and reframing metadata-dependent systems as pervasive, infrastructural 'metadata material discourse.' This undertaking has several useful implications for the politics of metadata (i.e. how metadata is regarded in the wider world, and how we as metadata professionals might engage the populace on the subject) as well as for professional practice and future work.

2. Historical Background

Metadata, a term familiar in libraries and other information-related fields, became known to information professionals in the 1990s as OCLC developed the Dublin Core metadata standard for describing objects. Around this time, a sister standard, the Resource Description Framework (RDF) was developed alongside Dublin Core so that both could be used for Semantic Web work, making the objects they describe even more accessible and interlinked (Harper, 2010). As Dublin Core and complementary encoding schemes like the eXtensible Markup Language (XML) and RDF gained steam, the nascent metadata community helmed by organizations like the Dublin Core Metadata Initiative (DCMI) had, by the late 2000s, created a conceptual identity around professional metadata that was separate from traditional cataloguing and classification, despite these other fields also dealing in descriptive work (Coyle, 2005). Henceforth, in the metadata community the term 'metadata' not only commonly referred to things (e.g. XML-encoded Dublin Core records or RDF serializations), but also to the practices of developing these specific kinds of things.

Metadata, however, would not remain a purely pragmatic thing, of interest to information professionals only. Within the last decade, metadata's political and ethical consequences have commanded the attention of a mass audience, most notably resulting from 'whistleblower' Edward Snowden releasing classified information about the assumedly private data—or better, metadata—that the US government collects on its citizens (Lyon, 2014). This revelation generated a tempest in the US and abroad, with average citizens asking, "What *is* metadata, and should I be worried about it?", as scholars and journalists scrambled to answer this complex question in blogs, articles, and books. Consequently, metadata has now assumed a primary place in our everyday concerns and activities, including (even if we are not conscious of it) in the ways we construct our love lives.

3. State of the Art

In light of the mainstream audience's interest in metadata and its relationship to surveillance, a Teen Vogue article defines metadata for its readers as not the 'content' of your photos and texts, but as the 'stuff' about them (Kobie, 2017). This simple definition geared at the ordinary populace is not a complete departure from the professional, reigning consensus on metadata. Definitions from the metadata community typically align with that of Marcia Zeng and Jian Qin's, with metadata described as "structured, encoded data that describe the characteristics of

information bearing entities and as such enable functions for identifying, discovering, assessing, and managing the entities" (Zeng & Qin, 2008). This definition encapsulates metadata work in its commonplace, practical sense. It is thus a useful pragmatic definition. (Coyle, 2005; Mitchell & Greenberg, 2009).

Metadata-dependent systems, for instance digital libraries and repositories, are often evaluated according to how well the implementation and usage match professional definitions and criteria like Zeng and Qin's, or relatedly, if metadata is correct according to prevailing standards (e.g., the Dublin Core). For example, Sarah Shreeves et al. and others have done studies on metadata usage in digital libraries, with a focus on how individual records do or do not flout rules of the Dublin Core (Shreeves et al., 2005, 2006). Analyses, like the ones Shreeves et al. perform of metadata dependent systems, are often rigorous in their focus on the empirical data a metadata record presents to us (e.g., does the dc:format element actually include a format in the value?). However, some scholars with an interest in information studies suggest other methods which might complement such empirical rigor. The late Claudio Ciborra writes of a scientized approach to systems analysis:

> [By] adopting the scientific mode of discourse, systems methodologies turn themselves away from everyday human dealings with technology, and find a (shaky) refuge in general and abstract dispositions and norms. They dislodge the problem of human existence out of the development and use of systems, and attempt to fill this ontological gap with the appearances of logic, objects, standards, and measurements, to, as concerned practitioners all over the world can testify, little avail (Ciborra, 2002).

Although studies like Shreeves et al.'s usefully and in great detail point to how metadata usage is often 'incorrect' in systems like digital libraries, and how this affects things like metadata harvesting, we might be left wondering about an implemented schema's sociopolitical or ethical consequences. With this in mind, we can follow Ciborra and perhaps reframe the professional notion around metadata as more complex than it seems at first glance. Contemplating what metadata as a discipline of practice should be, we can also reconsider how metadata, as a thing, manifests beyond the pragmatics of definitions and standards. We know that quantitative methodology has a useful place in assessing professional metadata practice. We have also addressed that an idea of metadata is now on the average person's radar, a 'lived world' phenomenon with potential numerous consequences. We can then ask: Can 'professional' metadata and 'popular' metadata meet in the middle? Can metadata practice 'learn' from the popular existence of metadata, that is, metadata in everyday, non-bibliographic platforms?

Such a consideration leads quite naturally to interrogating different philosophies and approaches to social science. The scientific method is one means to understand our world and the things in it (note that this method correlates with Ciborra's take on the *de facto* method of systems analysis), but many philosophers since the mid-20th century entertain *qualitative,* multiperspectival, highly interpretive conceptual views of everyday phenomena, including disciplines and their objects of study. Library and information science, and its constituent areas such as metadata, can also be read philosophically. Useful for this undertaking is *critical theory*, a philosophy first developed in the Frankfurt School in the 1930's which reframes everyday phenomena—and our relationships with them—using ideologies and other constructs (e.g., power, labor, identity, technology). This reframing—or *problematization*—changes a phenomenon taken for granted, or regarded commonsensically in our lived world, into a *critical* object of study. Problematizations in LIS have been posed by other researchers, addressing things like: 'Are bibliographic subjects objectively true, or subjectively constructed?', 'Does feminist theory help us more equitably understand web technology such as algorithms which judge traits like beauty, or schemes like Library of Congress Subject Headings?', 'Should we understand information retrieval in terms of labor or philosophy of language?', and so on (Blair, 1992; Furner, 2012; Rieder, 2016; Warner, 2010). Following this approach, we could complement Shreeves et al.'s study by asking, "What are the ethical, feminist, queer or political implications

of the Dublin Core application profiles (and element/attribute mismatches) implemented by particular organizations?"

Problematizations of metadata as a practice and a thing are somewhat uncommon within the metadata community—but that is not to say problematizations do not exist. For example, addressing an information science audience, Fidler and Acker go beyond ordinary definitions and understandings of metadata to problematize the Host-Host protocol ARPANET as metadata infrastructure providing a web of *infradata* (Fidler and Acker, 2017; we will revisit other, more pertinent *problematizations* of metadata imminently). On the other hand, popular culture outlets, including news media, are rife with metadata problematizations, or at least discussions of metadata's consequences. To this effect, 'metadata' has accrued problematic or negative connotations among lay consumers of information. The NSA-led collection of telephone and internet metadata revealed by Edward Snowden instigated privacy watchdog groups to monitor the pervasive reach of metadata, with one group stating "an individual's patterns of behaviour, viewpoints, interactions and associations" make it possible to "compile a very detailed and invasive picture of the entire population including their behaviours and interactions" (Privacy International, n.d.). Controversies raged about the extent to which this collection may or may not have violated the constitutionally-guaranteed privacy of citizens; while US President Barack Obama famously assured Americans that "nobody is listening to your telephone calls" (Obama, 2013), many argued that metadata itself tracks a significant amount of one's daily life. To quote Snowden himself, "'Metadata' means records about your private activities and associations. It's an activity dossier" (Snowden, 2015). Big data scientists have voiced concerns that "the questions raised... suggest that an ethical turn becomes more urgent as a mode of critique" (Lyon, 2014). In this vein, the scholar, journalist and privacy expert Zeynep Tufekci addresses metadata ethicality by telling Teen Vogue readers in a recent article what encrypted phone apps they should use for texting and sending photos to avoid government scrutiny. For a wider audience, Tufekci addresses surveilling potentials of metadata in her recent book on social media and worldwide protest movements (Chotiner, 2017). News media have begun to unpack the ways in which the reconstruction of a person's identity via their metadata (specifically, social media metadata) can have consequences in the immigration arena. For instance, US President Donald Trump recently announced intentions to screen select visitors to the United States via required handoff of social media usernames (Kravets, 2017).

What in particular compels us as authors of this paper are non-governmental uses of social media, namely profitable apps for dating, which critics have noted are pervasive in our society (Levine, 2015). We hope, through problematization, to demonstrate that the phenomenon of modern dating sites (with the particular example of the popular mobile app, Tinder) is wholly reliant on metadata from various sources that both monitors and constructs its users' virtual presences, sometimes to discriminatory effects.

4. Critical Occasion

Via problematization of Tinder, we aim to contribute to this changing idea of metadata in our own way, from within the metadata community. The opportunity now arises for us to explain our position around metadata. Both of us, the authors of this paper, studied library and information science and self-identify as 'metadata people.' One author (Deborah Maron) worked in metadata and digital libraries for several years, approaching the subject of metadata rigorously in an empirical sense; although evaluating accuracy in standards implementations in this manner is still important to her, she was made aware of critical theory as a useful apparatus for problematizing metadata due to recent pursuits in communication studies and philosophy. The other author (Erin Carter) is a technology professional with a background in communication and media studies who has also worked in metadata. We, the authors, share an interest in the effects of communication technology on society, and a concern for the social justice issues which arise from this interaction. Recent conversations around current events and metadata's nascent role in them, as well as our mutual vested interest in metadata overall, lead us to consider how a scientific

mindset could pair with other critical apparatus for doing the job of metadata or just examining metadata 'as a thing.' Additionally, we are millennials in the online dating milieu—specifically, we are white/female/feminist users of dating app Tinder.

Part of our perspective involves looking at different fields using critical theory. Particularly influential is Science and Technology studies (STS), a discipline which makes liberal use of critical theory at the intersection of humanity and technology. Recent scholarship in STS has problematized metadata in ways related to, for example, labor, capital and surveillance. Some individuals writing in STS come from library and information science (and their works are represented in publications like JASIST, Knowledge Organization, Library Juice, and Journal of Documentation, among others). Drawing from critical theory, STS, and our position, we aim to make two contributions in this paper:

1. Our general contribution: We hope to contribute to the longstanding tradition of STS scholars examining technology's role in the context of our evolving societal landscape, while also contributing to members of the metadata community's recent forays into problematizing metadata.
2. Our specific contribution: Following Lapôtre (2017), we contend that although oftentimes data and metadata are kept completely distinct, we believe that both are actually *material* (that is: *active*), and through a critical lens, the ways we study data can sometimes become the ways we study this 'material metadata discourse.' Further, we contend that there are alternative 'problematized' ways to read a system comprised of metadata that diverge from the positivist methods typically invoked, and these problematized readings can in turn shed light on other issues such as ethics, identity, surveillance, discrimination, credibility, etc. We hope in this foray to instigate critical discussion by treating more things perhaps typically considered 'data' only, as material and problematized metadata, an action which has particularly salient consequences for identity formation online.

We look to STS as well as constructs from literary theory, sociology, and Francophone philosophy to explore one facet of material metadata discourse in our lived experience: social media's 'metadata.' As previously mentioned, social media metadata is currently part of the cultural dialogue, and critical in identity building online. With social media in mind, we define a critical substrate involving the concepts of metadata-as-infrastructure (Jeff Pomerantz) and material discourse (from various rhetoricians and Digital Humanities scholars). From there we use différance (Jacques Derrida) and digital trace (from Bruno Latour and other Francophone philosophers) as methods to problematize social media metadata as a pervasive material infrastructure. Our case study and method involves a popular web resource—Tinder—with the user-facing information of the software not considered data, but instead, 'material metadata discourse.' We now offer a companion method to the positivist read of a system through our use of problematization and critical theory for the specific case of Tinder.

5. Conceptual Analysis

5.1. Metadata as Infrastructure and Material Discourse

Scholars such as Zeng and Qin contend that metadata is everywhere (2008). However, the evolution of 'the professional metadata account' or the way metadata has been professionalized has limited the scope of what they and others consider to be metadata in a theoretical and practical sense. For the purposes of this paper we favor a more liberal account derived from problematization. Another LIS scholar and metadata instructor—Jeff Pomerantz—in his recent book *Metadata,* uses an STS lens in problematizing and critiquing what he sees as the pervasive infrastructure (or "metadata grid") that undergirds a world of technology and human interaction:

> Metadata is infrastructural, like the electrical grid or the highway system. These pieces of modern infrastructure are indispensable but are also only the tip of the iceberg: when you flick on a lightswitch, for example, you are the end user of a large set of technologies and policies. Individually, these technologies and policies may be minor, and may seem

> trivial… but in the aggregate, they have far-reaching cultural and economic implications. And it's the same with metadata. Metadata, like the electrical grid and the highway system, fades into the background of everyday life, taken for granted as just part of what makes modern life run smoothly (Pomerantz, 2015).

A metadata electrical grid metaphor is not outlandish, given that scholars outside library science have relied on more radical metaphors to problematize pervasive technology, like Cthulu for describing biophysics and kinship, or hospitality industry standards for how we should 'treat' information and communication technology (Ciborra, 2002; Haraway, 2015). Pomerantz's explanation is useful to us because he contends metadata drives everyday digitally mediated experience, down to our ATM transactions—and by extension, the mobile apps we use to enhance our lived experience. It is precisely this pervasiveness that allows us to say that 'metadata can be anywhere or anything.'

Our second critical contention construes metadata studies as part of a tradition of examining *material discourse.* We firstly refer to 'discourse' in its more colloquial and basic sense, as spoken or written content that conveys meaning. Yet, "discourses are more than ways of thinking and producing meaning. They constitute the 'nature' of the body, unconscious and conscious mind and emotional life of the subjects they seek to govern" (Weedon, 1987, p. 108). Discourses are also a "form of *power* that circulates in the social field and can attach to strategies of domination as well as those of resistance" (Diamond and Quinby, 1988, as cited in Pinkus, 1996). Second, to take something as a *material* permutation is not merely to address its tangibility; materiality also accounts for a willingness/ability to engage, or how we engage, with that material thing. That is, material is *active*. To bring in an example from an LIS scholar, Johanna Drucker reconsiders computer infrastructures and things like algorithms as material in Digital Humanities scholarship (Drucker, 2013).

'Material discourse' typically transcends the textual realm, but when we take cues from literary theory, nearly any material discourse can be examined for qualities similar to those of text. Digital rhetoricians have reimagined things like forums, blogs, app platforms and much other "website stuff" as material discourse (Eyman, 2015). We propose that much web metadata exists materially; it is something with which we engage. This holds even more water if we bring in discussions of the nature of documents and information. Michael Buckland and Suzanne Briet claim non-textual things (even antelopes!) can hold informational value; therefore, it seems reasonable that metadata, as information things, can exist in various material permutations, as well (Buckland, 1997). Notably for us, Raphaëlle Lapôtre (2017) has recently problematized bibliographic metadata as a material phenomenon. We extend her contention slightly, proposing the concept of 'material metadata discourse,' which might challenge conventional and popular wisdom about the limits of metadata. We speak here of data which describes people, places and things (metadata); with which we interact (materially); and which also has consequences for power and ethics by monitoring and constructing us (discourse).

5.2. Case Study: Tinder as *Différance* and Digital Traces

Jacques Derrida originated the term *différance* as a two-fold concept describing the way meaning is constructed and transferred. Derrida says that the meaning of information is only intelligible in relation to other information which describes, defines, and informs it—thus, when information pertaining to a topic is accessed by a reader, its meaning is inevitably 'deferred' through a chain of other associated meanings which are called forth through the interaction between the reader and their knowledge of whatever other information has been used to describe, define, and inform that topic in that particular reader's experience (Derrida, 2002). We can apply Derrida's concept of *différance* to the study of 'digital traces,' as traces themselves are active *material*. Digital traces refer to the composition of of "a 'digital identity'... [as] the collection or the sum of digital traces—be they written, audio or video documents, logins, online purchases, or browsing sessions—that are left behind, deliberately or unconsciously, throughout the network of

a user's online relationships and exchanges" (Riegeluth, 2014). Digital traces are an unavoidable byproduct of computer-mediated human interaction. There is always a trace in text (the *différance*) of something, some implication or meaning, that came before in complex web interactions and interrelations. Metadata, understood in terms of digital trace/*différance*, is an amalgam of different temporalities and web loci (i.e., material metadata discourse) that describe and define an object of interest—in this case, a human being.

How is it that metadata affects our virtual and corporeal identity, and how do we allow it to do so when engaging with web technologies? We can understand digital traces as being constructed through a chain of other digital traces informing a user's persistent identity throughout various platforms and creating an overall digital depiction of that person—we can call this a 'digital avatar.' Other information scholars have studied digital identity formation in recent years (see Carter et al.'s discussion of information systems as "medium, determinant, and consequent of identity" [Carter et al., in press]). One place people build representations of themselves through digital traces is on a popular mobile dating app, a piece of material metadata discourse: Tinder.

Tinder does not allow account creation without a Facebook profile; a profile is generated via Facebook's (and other apps') *digital traces*, the traces themselves material metadata discourse. When a person creates a dating profile on Tinder, they grant permission for the app to view and use metadata that describes them from their Facebook profile. Facebook, which many of us know as a confluence of personal opinions, experiences, and other material evidence, forms a virtual idea of a 'corporeal body.' A person's Tinder profile, then, is partially constituted by Facebook's notion of one's past and present. Tinder is one place of many for the digital avatar to flourish, and, as a metadata locus, goes beyond quotidian corporeal attributes like age and location to present potential romantic partners with even more interesting information about a person.

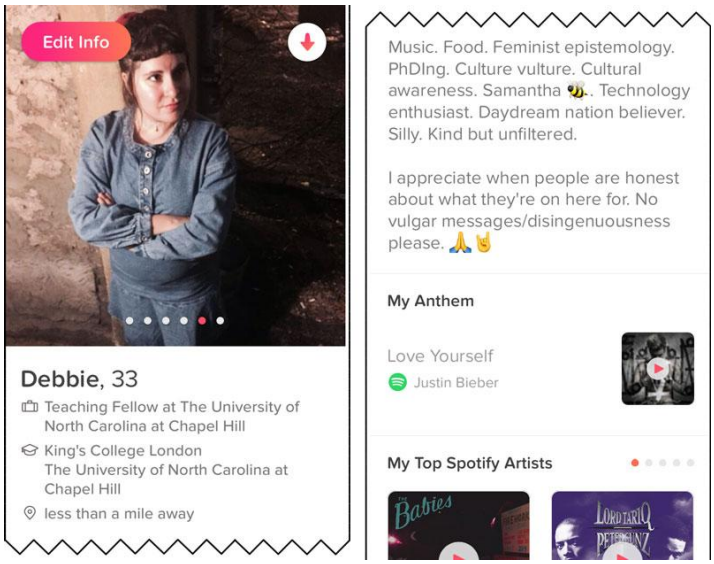


FIG. 1. Tinder profile illustrating an egregious default musical anthem, among other metadata

Metadata traces precipitate action and are themselves active material, fluid and engaged in discourse/narrative construction with one another. They also are construed here as 'meta' by virtue of their role in describing the virtual and corporeal self in digital dialogue. We define two major types of digital trace metadata in Tinder: *referential,* that which refers to the self or to other individuals in an obvious way on the interface, constituting the user-facing profile itself; and *transactional,* e.g. logs/statistics tracked by Tinder which are co-constructed by user behaviors and algorithmically determine matches for a user based on their perceived dating 'value' (note that this is an opaque process which a user knows little about, but its existence has been confirmed by Tinder). But what, exactly, constitutes a digital trace, or piece of metadata, in the

context of Tinder? Digital communication scholars have already established a piece of referential metadata—digital photographs—as 'digital traces' or 'digital footprints' which can be paired with an individual's online identity and traced through time and space (Girardin, 2008). Users choose their Tinder profile photographs; some are innocuous, while others depict wild nights out, complete with illicit drug use, drinking to excess, etc. Such life choices-as-referential-metadata trace from other social media such as Facebook.

Consequently, although Tinder profiles are on the surface *singular* entities representing a single person, digital traces can assign relative value to profiles, and profile metadata can transactionally affect the value of other people's profile metadata on Tinder. For instance, even if someone does not include a particular incriminating photo on Tinder, its existence can still be corroborated by that piece of 'material metadata' having been spotted on other dating sites. It is a trace which is, inescapably, part of a user's avatar that exists across digital space. The 'original' digital traces of a person's Facebook identity to which Tinder profiles refer are not even the person's authentic corporeal self, of course; rather, what one sees is self-created and self-selected metadata. A user has reconstructed a dating avatar (from metadata) using digital traces when they create a new social media profile which connects to extant social media accounts. The metadata describes a digitally curated (and questionably veracious) alternate version of the corporeal person, although the tendency to collect, select and project a more positive image has been well documented for more than a decade (Ellison, 2006).

In addition to adding metadata from a Facebook profile, Tinder users are invited to connect other social media accounts, such as Instagram and Spotify, which proffer more sources of referential metadata about a person in pictures and music tastes. When Spotify is connected, one's tastes, perhaps for particularly unfashionable musicians, are linked by default, although a person can alter this metadata to include artists and songs they would rather others see. Interestingly, when Instagram is connected, the user is offered much less of an opportunity to curate and customize their 'metadata traces' than they are with the initial granting of access to Facebook data. These tastes are inked indelibly onto their Tinder profile; there is no way to hide one's proclivity to egregiously photograph one's food (or take 'selfies,' for that matter). Just as Latour contends that other virtual centers of identity co-influence a particular locus or instance of identity (Latour, 2012), Instagram and Spotify traces likewise inform and influence prospective romantic matches on Tinder.

Other influential criteria that might attract or dissuade potential romantic partners in 'real life' are amplified in dating apps like Tinder, perpetuating various forms of discrimination: such criteria (in our scheme, translated in/as metadata presented to users either visually or textually, explicitly or implicitly) are ethnicity, class and age. Cultural theorists have written about these issues in the context of dating apps; Juana Rodriguez "argues that online space is an informational assemblage that reproduces colonial relations of power to construct racial otherness" (Rodriguez, as cited in Raj, 2011). Francisco verifies this in a recent article about Tinder, writing that "...black women and Asian men are the demographics on which the highest number of people swipe 'left,' thereby rejecting them... Black women and Asian men make up two demographics that have been long stigmatized as not-ideal sexual and romantic partners" (Francisco, 2017). Tinder also defaults to include metadata traces from Facebook such as one's education, occupation, and institutional affiliations; these traces inform the user of a potential partner's likely income bracket and perhaps, if such affiliations are elite, an elite 'valuation.' Additionally, Tinder's pricing structure reifies virtual social class and age: while most users use the free version of the app, a user can pay for a number of enhancements to their basic profile with a premium account, called 'Tinder Plus' (featuring the ability to have their profile shown to more people, etc.). This upgrade comes with an age-determined pricing structure: For those aged 29 and younger, the cost is ~$10 per month; however, starting at age 30, the price doubles (Abel, 2015). In this way, a user's digital traces have a direct financial impact on their pursuit of love, and simultaneously devalue them as they age. In sum, these traces of ethnicity, class and age appear initially as materially-discursive, *referential* metadata in profiles (names, pictures and otherwise), but translate into *transactional*, 'behind the scenes' value when other users utilize

them to make a determination of interest or lack thereof (and when Tinder monitors such metadata to determine who are the most 'swiped on,' 'valuable' users to match with other 'valuable' users).

6. Discussion

Consideration of metadata as a pervasive infrastructural and material phenomenon, in the way we have construed it here, has a myriad of personal and political consequences which are relevant to Science and Technology Studies' frequent focus on ethical concerns. Our digital traces, which we consider a way to perceive aspects of material metadata discourse, affect how we as *users* build identity and are surveilled, as well as how others treat us. We consider Tinder not just a system to be read as a 'bunch of metadata' with no intended effects, but as a system of active material discourse that constructs the user and their experience in that system. It has been documented anecdotally and via data collection and analysis that Tinder perpetuates a cycle of inequity through their use of material metadata, "bumping" certain users up and demoting others based on ethnicity (Francisco, 2017). Likewise, library science has, for many years, maintained an interest in the effects and unequal distribution of library utilization and information-seeking behavior of minority groups (Spink and Cole, 2001). Notably, African-Americans appear to be inadequately served in both Tinder and library systems (Hughes-Hassell, Bracy and Rawson, 2017).

How should we as metadata professionals critically consider the boons and banes of our surveilling approaches and perhaps communicate our techniques with a metadata-focused populace? How does surveilling in the Tinder sphere cast back on how we view gender, social and racial equitability in use of systems such as libraries? We consider two examples for the professional practice of workplace metadata technology which could benefit from a critical treatment like the one presented here. In the library context, we need only to look to transaction logs to find user-generated metadata which can be used for surveillance or monitoring purposes. Although library patrons may be less aware that they are creating metadata, they do so as they select and borrow materials. Should we as metadata professionals think about how the metadata driving our systems might in some cases perpetuate inequity, or that perhaps we have not been 'keeping an eye out' for inequitable systems? As *professionals*, if we do not consider metadata implications as opportunities for study by metadata professionals, we risk losing out on not only improving our systems but also a larger cultural moment and an opportunity to share our knowledge and experience with an eager wider audience. Claudio Celis addresses a communication studies audience when he argues that we live in "societies of metadata" resulting from a cyber, machinic and human mashing; so why don't we as metadata professionals heed the call to pursue the effects of such societies (Celis, 2015; Fidler & Acker, 2017)? In assimilating concerns such as those learned from Tinder, we need not abandon the rigorous approaches we use to examine things like metadata standards implementations, but instead we can complement such approaches by turning a critical eye to the systems implementing metadata by examining additional ethical issues. Critical theory (our choice, but by no means the only choice), coupled with something like digital rhetoric's ideas around material discourse, allows us as both researchers and practitioners of metadata to remain skeptical, critical and multiperspectival in the work we produce and consume.

Another example appears in the world of bibliographic metadata: the digital library. So many LIS systems, like digital libraries, are metadata dependent. Considering that many systems are also being RDF-ized, digital libraries—as records or entire archives—are ever more transportable, shareable, interoperable, engageable and active with other points of metadata; that is, they are *materially discursive metadata* as well as webs of *digital traces.* In the manner we investigate Tinder, examining its implications as a digital trace mechanism that designs identity and inequity, we can also look at the ethical implications of material metadata discourse such as digital libraries by considering that the encoded content itself has *reifying* potential if it contains

erroneous information about people, places, or things (Thomas, 2012). Such erroneous metadata can, at times, neither be changed nor ignored by people interacting with it.

7. Conclusion

By considering more things to be materially discursive metadata, and considering how the wider world's conceptions of metadata are consequential for our professional endeavors (and of course, how wider world conceptions cast back onto our profession), we gain an opportunity to extend the scope of what we do while improving things from 'inside' the metadata world. In this paper, we examine online identity formation via 'material metadata' in social media and the dating app Tinder. This effort 1) contributes to Science and Technology Studies as well as Metadata Studies by promoting critical theory and problematization from inside the metadata community; and 2) provides complementary methods—specifically, the apparatus *différance and digital trace*—to scientized methods for problematizing and reframing metadata-dependent systems as pervasive, infrastructural 'metadata material discourse,' an act which also blurs the distinction between data and metadata and so opens doors for research by our community.

If practitioners and scholars of metadata deem critical theory a useful companion methodology to scientific rigor, we might imagine a rich possible future of research opportunities. Using referential and transactional digital traces, for instance, we could explore the consequences of AirBnB-as-metadata, or Uber-as-metadata. One other critical concept we believe holds much promise is the *boundary object*, something that retains its integrity across spaces while adapting to different communities' needs (e.g. a map, to a cartographer, is something different than a map, to a museum collector—yet it still remains a 'map'; thus, it is a boundary object) (Bowker & Star, 1999). We believe boundary objects could help frame and criticize things like metadata crosswalks, or other newly-emerging forms of 'material metadata discourse' such as online forums, memes, and other materials which take different identities across the communities that repost and comment on them.

But we need not rest only on structural metaphors in critical theory. We can also examine something like a professional metadata creator's intentions, as well as ramifications for users of metadata, by looking at *epistemologies* illuminating the positions, motives and beliefs of individuals and institutions. For example, feminist or Marxist epistemologies can frame Dublin Core or customized Application Profile usage to help us understand an organization's values or how an implementation might contribute to resistance movements, etc. (Pastva, 2014). Critical theory is but one item in the critical toolbox. Critical theory-as-method can help us successfully breach multiple worlds in a way that enables metadata to be 'more than what it seems.'

Acknowledgements

We would like to thank the organizational reading group (org) as well as our DCMI conference reviewers for their invaluable comments on earlier iterations of this paper.

References

Abel, J. (2017, February 21). Tinder Plus costs $10 monthly unless you're 30 or older; then it's $20. *Consumer Affairs*. Retrieved September 16, 2017, from https://www.consumeraffairs.com/news/tinder-plus-costs-10-monthly-unless-youre-30-or-older-then-its-20-030615.html

Blair, D. C. (1992). Information retrieval and the philosophy of language. *Computer Journal*, 35, 200–207.

Bowker, G., & Star, S. L. (1999). *Sorting things out*. Classification and its consequences.

Buckland, M. K. (1997). What is a "document"? *Journal of the American Society for Information Science (1986-1998)*, 48(9), 804.

Carter, M., Compeau, D., Kennedy, M., & Schmalz, M. (In Press). The Content and Context of Identity in a Digital Society. Proceedings 25th European Conference on Information Systems (ECIS 2017). Paper to be presented at 25th European Conference on Information Systems, Guimarães, Portugal, 5–10 June 2017.

Celis, C. (2015). The Machinic Temporality of Metadata. *tripleC: Communication, Capitalism & Critique*. Open Access Journal for a Global Sustainable Information Society, 13(1), 101–111.

Chotiner, I. (2017, May 08). Has Protesting Become Too Easy? *Slate*. Retrieved June 10, 2017, from http://www.slate.com/articles/news_and_politics/interrogation/2017/05/zeynep_tufekci_author_of_twitter_and_tear_gas_on_networked_protest.html

Ciborra, C. (2002). *The Labyrinths of Information: Challenging the Wisdom of Systems*. OUP Oxford.

Coyle, K. (2005). Understanding metadata and its purpose. *The Journal of Academic Librarianship*, 31(2), 160-163.

Derrida, J. (2002). *Différance* (p. 24). Na.

Drucker, J. (2013). Performative materiality and theoretical approaches to interface. *Digital Humanities Quarterly*, 7(1).

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2), 415-441.

Eyman, D. (2015). *Digital rhetoric: Theory, method, practice*. University of Michigan Press.

Fidler, B., & Acker, A. (2017). Metadata, infrastructure, and computer-mediated communication in historical perspective. *Journal of the Association for Information Science and Technology*, 68(2), 412–422.

Francisco, E. (2017, September 13) How Tinder Exposed Our Reliance on Racist Stereotypes. *Inverse Culture Magazine*. Retrieved September 16, 2017 from, http://www.inverse.com/article/36379-tinder-black-women-asian-men-racism

Furner, J. (2012). FRSAD and the Ontology of Subjects of Works. *Cataloging & Classification Quarterly*, 50(February 2014), 494–516.

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4).

Haraway, D. (2015). Anthropocene, Capitalocene, Plantationocene, Chthulucene: Making Kin. *Environmental Humanities*, 6(1), 159–165.

Harper, C. (2010). Dublin Core Metadata Initiative: beyond the element list. *Information Standards Quarterly*, 22(1), 20-28.

Hughes-Hassell, S., Bracy, P. B., & Rawson, C. H. (2017). Libraries, literacy, and African American youth: research and practice. Santa Barbara, CA: Libraries Unlimited.

Kobie, N. (2017, March 2). How to keep messages secure. *Teen Vogue.* Retrieved June 04, 2017 from http://www.teenvogue.com/story/how-to-keep-messages-secure

Kravets, D. (2017, June 2). Trump administration rolls out social media vetting of visa applicants. *Ars Technica.* Retrieved June 04, 2017, from https://arstechnica.com/tech-policy/2017/06/trump-administration-rolls-out-social-media-vetting-of-visa-applicants/

Lapôtre, R. (2017). Library Metadata on the web: the example of data.bnf.fr. *JLIS.it*, 8(3), 58-70.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). 'The whole is always smaller than its parts'–a digital test of Gabriel Tardes' monads. *The British journal of sociology*, 63(4), 590-615.

Levine, D. (2015, February 12). Online dating - the psychology (and reality). *Elsevier Connect.* Retrieved June 04, 2017, from https://www.elsevier.com/connect/online-dating-the-psychology-and-reality

Lyon, D. (2014). Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data & Society*, *1*(2), 2053951714541861

Mitchell, E., & Greenberg, J. (2009). *Metadata literacy [electronic resource] : an analysis of metadata awareness in college students*. University of North Carolina at Chapel Hill, Chapel Hill, N.C. Retrieved from http://search.lib.unc.edu/search?R=UNCb6351578

Obama, B. (2013, June 7). Statement by the President. *White House Briefing Room Statements & Releases.* Retrieved June 04, 2017, from https://obamawhitehouse.archives.gov/the-press-office/2013/06/07/statement-president

Pastva, J. and Harris, V. (2014, September 9). PunkCore: Developing an Application Profile for the Culture of Punk. Poster session presented at *International Conference on Dublin Core and Metadata Applications*, Austin, TX.

Pinkus, J. (1996, August). Foucault. Retrieved June 10, 2017, from http://www.massey.ac.nz/~alock/theory/foucault.htm

Pomerantz, J. (2015). *Metadata*. MIT Press.

Privacy International. (n.d.). Retrieved June 04, 2017, from https://www.privacyinternational.org/node/53

Raj, S. (2011). Grindring bodies: Racial and affective economies of online queer desire. Critical Race and Whiteness Studies, 7(2), 1-12.

Rieder, B. (2016). Scrutinizing an algorithmic technique: the Bayes classifier as interested reading of reality. *Information, Communication and Society*, 4462(June), 1–18.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, April 7–10, 2005. Association of College and Research Libraries, 223.

Shreeves, S. L., Riley, J., & Milewicz, L. (2006). Moving towards shareable metadata. *First Monday*, 11(8). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1386/1304

Snowden, E. (2015, November 02). "Metadata" means records about your private activities and associations. It's an activity dossier. The novelty is in the lack of warrants. Retrieved June 04, 2017, from https://twitter.com/snowden/status/661302234068701188?lang=en

Spink, A., & Cole, C. (2001). Information and poverty: Information-seeking channels used by African American low-income households. Library & information science research, 23(1), 45-65.Thomas, N. (2012). *Social Computing as Social Rationality.*

Warner, J. (2010). *Human information retrieval*. MIT Press Cambridge, MA.

Zeng, M. L., & Qin, J. (2008). *Metadata*. Neal-Schuman Publishers.

Zervas, G., Proserpio, D., & Byers, J. W. (2014). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research*.

IFLA LRM – Finally Here

Maja Žumer
University of Ljubljana, Slovenia
maja.zumer@ff.uni-lj.si

Pat Riva
Concordia University Library, Montreal, Canada
pat.riva@concordia.ca

Abstract

The IFLA Library Reference Model (IFLA LRM) consolidates the three models of the FRBR Family. In this paper, first the differences between the three models are presented as well as the major modelling and presentation issues identified. The main part is the general description of IFLA LRM. Only the most important features are presented, with examples illustrating the modelling approaches.

Keywords: IFLA Library Reference Model, IFLA LRM, conceptual models, FRBR

1. Introduction

With the publication of the *Functional Requirements for Bibliographic Records* (FRBR, 1998) the library community made a leap forward; this first conceptual model of the bibliographic universe provided the necessary foundation for the development of new generations of bibliographic information systems such as library catalogues and bibliographies. The two models that followed further developed the area of authority control: the *Functional Requirements for Authority Data* (FRAD, 2009) focused on agents and works, while the *Functional Requirements for Subject Authority Data* (FRSAD, 2010) developed the subject relationship. The three models, now usually referred to as the FRBR Family of models, were developed over time by different working groups and so some differences in structure and conceptualization between the models were not unexpected. While the differences are mostly in the details, when the three FR models were declared as element sets in the IFLA domain (http://iflastandards.info/ns/), the differences resulted in three different namespaces reflecting incompatible definitions of (some) entities, attributes and relationships. These differences constituted a definite barrier to successful and compatible implementations. The decision to harmonize the three models was therefore the logical next step. In 2011, the FRBR Review Group, the body within IFLA responsible for the development and promotion of the FRBR Family of models, started the consolidation process. In 2013, the Consolidation Editorial Group (CEG) was established and assigned the task of combining the three models into a consistent whole and thus preparing a unified view of the bibliographic universe.

LRM incorporates many insights gained through almost twenty years of experience in working with the FRBR Family of models. The development of LRM was viewed as an opportunity to critically assess the existing conceptual models, to identify and fill gaps, and to answer recurring questions. LRM maintained the aspects of the FRBR model that were validated through research carried out on end-user mental models (see Pisanski & Žumer, 2010). LRM offers a single complete model covering all aspects of bibliographic information, removing barriers to adoption due to divergent conceptualizations between the models of the FRBR Family. The approach was to clarify where experience indicated it was needed, such as in the modelling of aggregates and serials, and generally provide a more robust, rigorous model. Additionally, the presentation of the LRM model definition was designed so as to include information needed for the declaration in RDF of an element set reflecting the model. The aspects of the model that make it semantic web-ready were the topic of a recent IFLA congress presentation by the authors (see Riva & Žumer, 2017).

Already in November 2016, the RDA Steering Committee (RSC) adopted LRM, instead of the FRBR Family of models, as its conceptual model to underlie the current major development of *RDA: Resource Description and Access* (see http://www.rda-rsc.org/ImplementationLRMinRDA). As RDA is rapidly becoming the most widely used standard for description and access to bibliographic resources, this decision will have major impact in the bibliographic community with far-reaching ramifications. RDA will be the first large-scale application of the LRM model and will be a de facto extension of the model to cover all elements required in resource description. The confidence shown by the RSC in the approach taken in LRM is a strong validation of the consolidation process.

2. Major Differences between FRBR, FRAD and FRSAD

2.1. Style and Focus

While all three models use the same entity-relationship formalism, their presentation is rather different. The specification of FRBR is mostly written using free text and there are no strict boundaries between the definitions and the scope notes; there are also many examples, but they are not specifically explained. As seen from the many discussions following its publication, the components of the model are not always strictly defined and are open to interpretation. While this adds to flexibility, it hinders the interoperability of the systems developed using the model.

FRAD, on the other hand, is already more formal. Tables are used to define the model, but there is no clear boundary between the definitions, the scope notes and the examples, the latter mostly seem to be a part of the scope note. The specification of the relationships is more formalized, but still not presented quite systematically.

The structure of FRSAD, on the other hand, is very simple. Definitions and examples are clearly delimited, and although most attributes have few scope notes, there are no particular issues in interpreting the model.

As to the scope of the models, some differences can be noticed as well. FRBR and FRSAD are primarily end-user focused, which is obvious from the user tasks declared in FRBR: *find, identify, select, obtain*. In FRSAD *explore* is added, while *obtain* is not relevant. FRAD, on the other hand, is to some extent also modelling the cataloging process and, along with *find* and *identify*, introduces two additional tasks, *justify* and *contextualize*, which describe the work of a cataloguer performing authority control.

2.2. Entities

All three models keep the central entities (*work, expression, manifestation, item,* often referred to as WEMI) and their definitions are essentially unmodified. To FRBR's two entities defined to participate in responsibility relationships, *person* and *corporate body*, FRAD adds a new entity, *family*, and changes the definition of *person* to include "a persona or identity established or adopted by an individual or group". There are also major differences in the treatment of appellations: FRBR models them as mere attributes, while FRAD and FRSAD introduce specific entities to enable assigning attributes to an appellation itself. In contrast to FRSAD, which only defines one entity (*nomen*), FRAD defines three for different types of appellations: *name, controlled access point,* and *identifier*. In keeping with its scope, FRAD introduced two further entities, *agency* and *rules*, used in modelling the cataloguer's process in assigning *controlled access points*.

2.3. Attributes

In FRBR numerous attributes are defined for the four entities of the first group (*work, expression, manifestation, item*). These were drawn from an examination of the data elements typically included in bibliographic records formulated following ISBD, the IFLA *Guidelines for Authority and Reference Entries*, the IFLA *Guidelines for Subject Authority and Reference Entries*, and the *UNIMARC Manual*, although at a lesser degree of granularity. Using these sources led to the

inclusion of many specialized material-specific attributes, particularly for *expressions* and *manifestations*. The entity to which a data element should be attached was not always clear, with the result that the attribute *medium of performance* was considered both a *work* and an *expression* attribute, and some attachments were later disputed (such as expected regularity and frequency of serials). The close parallels between the attributes and the data elements led to uncertainty between the respective roles of the ISBD and FRBR. FRAD did not list all of the previously defined attributes, but did add certain *work* attributes typically recorded only in authority records. FRAD concentrated on expanding the attributes for *person* and *corporate body*, which were only minimally developed in FRBR, and also on proposing the attributes for *name* and *controlled access point*. FRSAD identified a very similar list of attributes for the *nomen* entity, but only the attributes *type* and *scope note* for *thema*.

2.4. Relationships

In FRBR and FRAD a distinction is made between the "primary" relationships, which are presented in the respective high-level diagrams, and all other relationships of interest, which are presented in tables. Cardinality is indicated in the diagrams illustrating the primary relationships, but is not given explicitly for any of the other relationships. In FRBR only additional relationships among WEMI are defined. Some of the same relationships appear in the tables for work-to-work, expression-to-expression (of different works), and expression-to-work relationships.

FRAD presents the additional relationships in three groups: those between *persons, families, corporate bodies* and *works* (this section actually covers WEMI, not only *works*), those between various *names* of *persons, families, corporate bodies* and *works*, and finally, those between *controlled access points*. The presentation of relationships in both FRBR and FRAD at times obscures the intended domain or range of the relationship. An example of this is in the subject relationship in FRBR, where the range is only indicated in a diagram by a box that encompasses all the entities declared in the model. However, this box is not itself named or identified as an actual entity in the model. In FRAD, the relationships are given a term (such as pseudonymous relationship, membership relationship) but not relationship names or inverse names. In each of these models, there are multiple tables of relationships, with no single comprehensive listing of all relationships.

2.5. Summary of Major Differences

The top five differences among the three models in the FRBR Family, in terms of their impact on the semantics of the models, are the following.

- User tasks

  Find, identify, select and *obtain* are defined by FRBR. FRSAD adds *explore*, intended to cover browsing and, consequently, serendipitous discovery. FRAD, on the other hand, focuses more on the cataloguing process and defines *justify* and *contextualize*.

- Definition of the *person* entity

  In FRBR, the entity *person* is a defined as "an individual", while in FRAD it includes also a "persona or identity established or adopted by an individual or group".

- Treatment of appellations as attributes or as entities

  In FRBR appellations are modelled as attributes of entities, in contrast FRAD and FRSAD introduce appellations as entities. While FRSAD defines only one appellation entity, *nomen*, FRAD lists three distinct entities for different types of appellations: *name, identifier* and *controlled access point*.

- Treatment of subjects as an attribute or a relationship

FRBR and FRSAD both define the *has as subject* relationship with the entity *work* as its domain, while in FRAD subject is modelled as an attribute of *work*.

- Relationships

  Relationships are modelled at different levels of specificity and, particularly in FRBR and FRAD, are not all declared in both directions and cardinality is not always specified. The domains and ranges indicated for some relationships do not indicate specific entities.

A detailed examination and comparison of the three models in the FRBR Family of models reveals many other points of divergence. In the *Transition Mappings* document, the FRBR, FRAD and FRSAD models are aligned where possible, and the mapping of each user task, entity, attribute and relationship declared in them with LRM is presented in full. This exercise also highlights all of the differences among the FRBR Family models.

3. IFLA Library Reference Model (IFLA LRM)

The task of the CEG was to:

- Prepare a high-level abstract model
- Use the entity-relationship formalism
- Develop a consistent model consolidating all three models of the FRBR Family
- Consider implementation in the Semantic web

The resulting model is described as (LRM, p. 6):

> The conceptual model as declared in IFLA LRM is a high-level conceptual model and as such is intended as a guide or basis on which to formulate cataloguing rules and implement bibliographic systems. Any practical application will need to determine an appropriate level of precision, requiring either expansion within the context of the model, or possibly some omissions. However, for an implementation to be viewed as a faithful implementation of the model, the basic structure of the entities and the relationships among them (including the cardinality constraints), and the attachment of those attributes implemented, needs to be respected.

3.1. User tasks

In line with the FRBR Family, in LRM the user tasks are central and form the starting point for model development. The tasks which need to be enabled and supported by a bibliographic information system define the scope of the model and are starting points from which the entities, attributes and relationships are declared. Bibliographic information systems are of interest to varied target audiences, from library users (readers, researchers, students…) to librarians and other actors in the information chain, including publishers and booksellers. These user groups have different needs and different priorities. LRM therefore follows FRBR in defining end-users, and librarians looking for information on behalf of end-users, as its primary audience. Librarians who create and maintain metadata may perform these same tasks as part of their work – they are included in this sense. On the other hand, the model does not include administrative data, which is otherwise essential for library operations, such as preservation or acquisitions metadata.

Five basic user tasks are defined (Table 1). The definitions are phrased by specifying the user's goal when performing each action. The term 'resource' is used in its broadest meaning, standing for any entity defined in the model. The tasks are listed in the order in which they are normally executed, which does not mean that they must all be performed each time a end-user accesses a bibliographic information system or that they cannot be repeated. Particularly *identify* and *select* often occur simultaneously and in interaction.

Table 1:	**User Tasks Summary**
Find	To bring together information about one or more resources of interest by searching on any

Table 1:	User Tasks Summary
	relevant criteria
Identify	To clearly understand the nature of the resources found and to distinguish between similar resources
Select	To determine the suitability of the resources found, and to be enabled to either accept or reject specific resources
Obtain	To access the content of the resource
Explore	To discover resources using the relationships between them and thus place the resources in a context

The first four tasks are essentially the same as the tasks with the corresponding names in FRBR, while *explore* was first introduced by FRSAD. The need for 'navigation' is already mentioned in FRBR and in subsequent years many researchers emphasized that a modern bibliographic information system needs to support browsing and, consequently, serendipitous discovery.

3.2. Entities

The CEG compared the definitions of entities across the three models and identified the semantically identical ones (such as *work, expression, manifestation, item*), the similar ones (FRAD *name* and FRSAD *nomen*) and the very different ones (*person* in FRBR and FRAD). All entities were critically reviewed and evaluated. The decision was to keep only the entities which were required due to having specific attributes or being used in specific relationships. In contrast with the FRBR Family, where all entities are at the same level, a hierarchical structure of entities is introduced in LRM by declaring entities within a structure of superclasses and subclasses. That one entity is a subclass of another entity can be expressed using the isA relationship. This powerful mechanism enables considerable simplification of the model, because attributes and relationships can be declared on the higher level and do not have to be repeated on lower levels.

Entities of the first group (often also called WEMI) remain basically the same conceptually; however, there are some changes in the wording of their definitions.

Table 2: Work, expression, manifestation, item

Work	The intellectual or artistic content of a distinct creation
Expression	A distinct combination of signs conveying intellectual or artistic content
Manifestation	A set of all carriers that are assumed to share the same characteristics as to intellectual or artistic content and aspects of physical form. That set is defined by both the overall content and the production plan for its carrier or carriers.
Item	An object or objects carrying signs intended to convey intellectual or artistic content

On the other hand, some changes were introduced in the second group by declaring a superclass, *agent*, and subsuming both *corporate body* and *family* into a broader entity termed *collective agent*.

Table 3: Agents

Agent	An entity capable of deliberate actions, of being granted rights, and of being held accountable for its actions
Person	Individual human being
Collective agent	A gathering or organization of persons bearing a particular name and capable of acting as a unit

The "agent" entities can best be presented using the basic relationships (Figure 1).

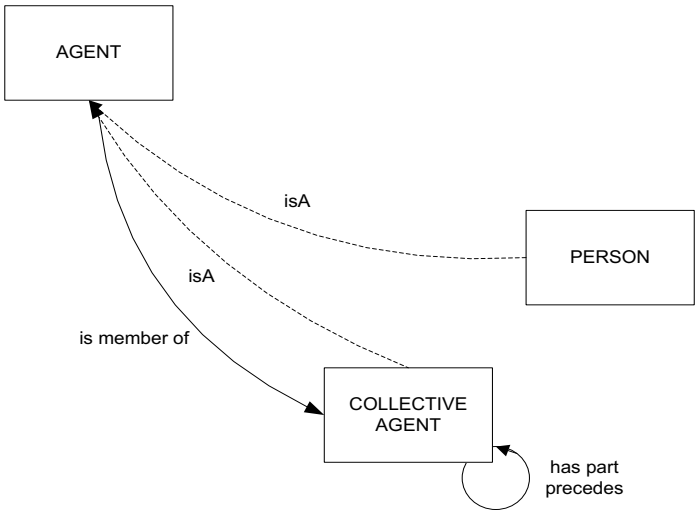


Figure 1: Agent relationships

In LRM, the entity *person* includes only living individuals or those who are assumed to have lived. Figures generally considered fictional, literary or purely legendary are not instances of the entity *person*. They can act as subjects of works. When they seem to be creators, it is in fact a *person* or a *collective agent* using that particular appellation in the context of that act. The name used does not change the nature of the agent. This follows the definition of the *person* entity in FRBR, and is unlike the FRAD approach which conflates real persons with bibliographic identities.

The FRSAD model first introduced the entities *thema* and *nomen* as the mechanism for modelling the appellation relationship. Both entities remain in LRM, with a slight label change; we have the term *res* replacing *thema* to avoid the restriction to the subject relationship implied in the term *thema*. *Res* is, therefore, the superclass of all LRM entities and *nomen* is the appellation used to refer to an instance of *res*. Modelling appellations as entities allows us to assign them attributes such as language, alphabet or controlled vocabulary and to establish relationships between different appellations for the same entity such as between former and later name of a *person*.

Table 4: Res and nomen

Res	Any entity in the universe of discourse
Nomen	An association between an entity and a designation that refers to it

Two new entities were added in order to model in more detail the spatial and temporal aspects: *time-span* and *place*. Using these entities, many characteristics previously modelled as attributes can be modelled as relationships in LRM.

Table 5: Place and time-span

Place	A given extent of space
Time-span	A temporal extent having a beginning, an end and a duration

3.3. Attributes

Attributes provide a mechanism for assigning particular characteristics to instances of entities. While FRBR and FRAD declare an exhaustive list of attributes applicable to particular types of entities, the decision was made to include only the most general and common attributes in LRM. The list of attributes is therefore not a complete inventory of characteristics that might be of interest and none of the attributes are mandatory. An application can define additional attributes to record additional relevant data or to record data at a greater level of granularity than is illustrated. Certain attributes that are important to the model or are frequently relevant in bibliographic systems are included here. However, the listing of an attribute in the model is not intended in any way to imply that these attributes are required for any application.

As illustration, the six attributes of *manifestation* are listed in Table 6. This contrasts sharply with the 38 *manifestation* attributes defined in FRBR.

Table 6: Attributes of *manifestation*

Category of carrier	A type of material to which all physical carriers of the *manifestation* are assumed to belong
Extent	A quantification of the extent observed on a physical carrier of the *manifestation* and assumed to be observable on all other physical carriers of the *manifestation* as well
Intended audience	A class of users for which the physical carriers of the *manifestation* are intended
Manifestation statement	A statement appearing in exemplars of the *manifestation* and deemed to be significant for users to understand how the resource represents itself
Access	Information as to how any of the carriers of the *manifestation* are likely to be obtained
Use rights	A class of use and/or access restrictions to which all carriers of the *manifestation* are assumed to be submitted

The *category of carrier* attribute is a sub-type of the *category* attribute defined for the entity *res*. Since *category* is defined for the top entity *res*, *category* attributes can automatically apply to any entity, whether declared for that entity or not. Despite this, the *category of carrier* attribute is one of the sub-types of the higher-level attribute that is explicitly declared in the model. This serves to illustrate some of the ways categorization can be used to record significant characteristics of entities, and to draw attention to the way LRM models certain FRBR *manifestation* attributes. The only other attribute of *res* is the *note* attribute, which automatically extends to all the subclasses of *res*, including *manifestation*, even though it is not explicitly declared.

The new attribute *manifestation statement* is a generalization of many FRBR *manifestation* attributes, particularly those drawn from ISBD. Any attribute that consisted of a transcription of a statement found in exemplars of a *manifestation* is actually a sub-type of this new general attribute. Transcription distinguishes a *manifestation statement* from a free-text or cataloguer-composed note, and is something that is specific to the *manifestation* entity. Defining this attribute at this functional level illustrates a mechanism in LRM that makes the model flexible and independent of any specific implementation. LRM does not prescribe the types of *manifestation statements* of interest. The application can sub-type this attribute to the level of granularity that suits the needs of the implementation.

Another significant generalization in LRM relates to the *work* attribute *representative expression attributes*. This is defined as "An attribute which is deemed essential in characterizing the *work* and whose values are taken from a representative or canonical *expression* of the *work*". This approach resolves the apparent contradiction between the assignment of certain attributes to the *expression* entity (such as *language, key, medium of performance, cartographic scale*) and the impression that values of these attributes are significant in delimiting the boundaries of the *work*. LRM follows the FRBR Family models in not labeling any particular *expression* as more significant and just allowing for the specification of a network of derivative relationships among *expressions*

of the same *work*. However, end-users do in some way consider certain *expressions* to more fully represent the "intent" of the *work*. The *expressions* that are viewed as most canonical or representative are often the original *expression*; however, due to the complexity of derivation networks, this is not always the case. It is the values of certain *expression* attributes that are seen in these representative *expressions* that are parked with the *work* entity via the *representative expression attributes*. Again, LRM allows each implementation to determine which *expression* attributes will function to characterize the *work*. The choice of attributes may depend on the form of the *work*.

3.4. Relationships

Relationships are the core of the model – they link entities and place them in context. Some of the relationships, for example the so-called primary relationships in FRBR, remain virtually unchanged in LRM, others differ primarily in the level of generality. LRM relationships are high-level and general, but they provide a framework for consistent extensions. Any entity can be linked to the entities *place* and *time-span* via the specific distinct association relationships (*res* is associated with *place* and *res* is associated with *time-span*). All relationships are refinements of the top-level relationship (*res* is associated with *res*). When needed, implementers can therefore add more granular relationships.

The relationships between *works, expressions, manifestations* and *items* are the center of the model and are in essence required. Relationships in general enable and support exploration and should be included as much as possible in implementations.

In LRM all relationships are declared specifying their domain and range, as well as the cardinality. Inverse names are also stated systematically.

Relationships between WEMI are illustrated in Figure 2.

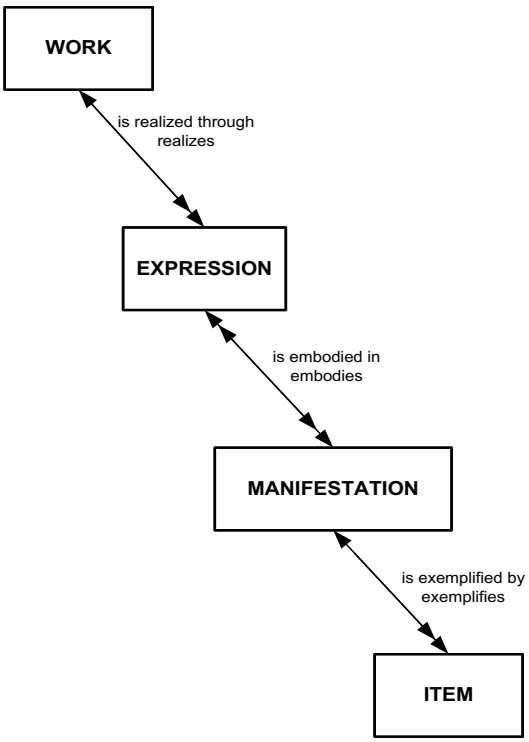


Figure 2: WEMI relationships

As an example, all work-to-work relationships are listed in Table 7.

Table 7: Work-to-work relationships

Domain	Relationship name	Inverse name	Range	Cardinality
Work	has part	is part of	Work	M to M
Work	precedes	Succeeds	Work	M to M
Work	accompanies / complements	is accompanied / complemented by	Work	M to M
Work	is inspiration for	is inspired by	Work	M to M
Work	is a transformation of	was transformed into	Work	M to 1

Agent-to-WEMI relationships have been streamlined as well (Figure 3).

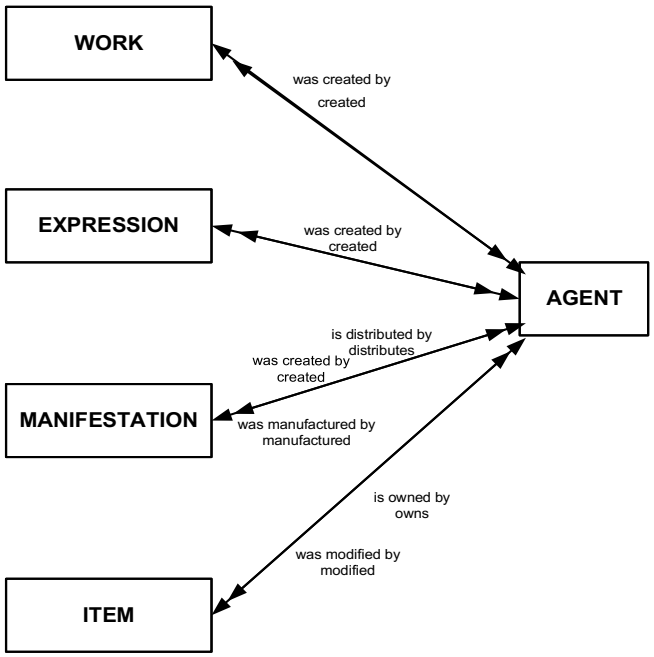


Figure 3: Agent-to-WEMI relationships

LRM declares 36 distinct relationships, as well as the relevant inverse relationships. The overview of all LRM relationships is shown in Figure 4. The isA relationships between all other entities and the entity *res* is not shown. For the sake of simplicity, relationships are shown in one direction only.

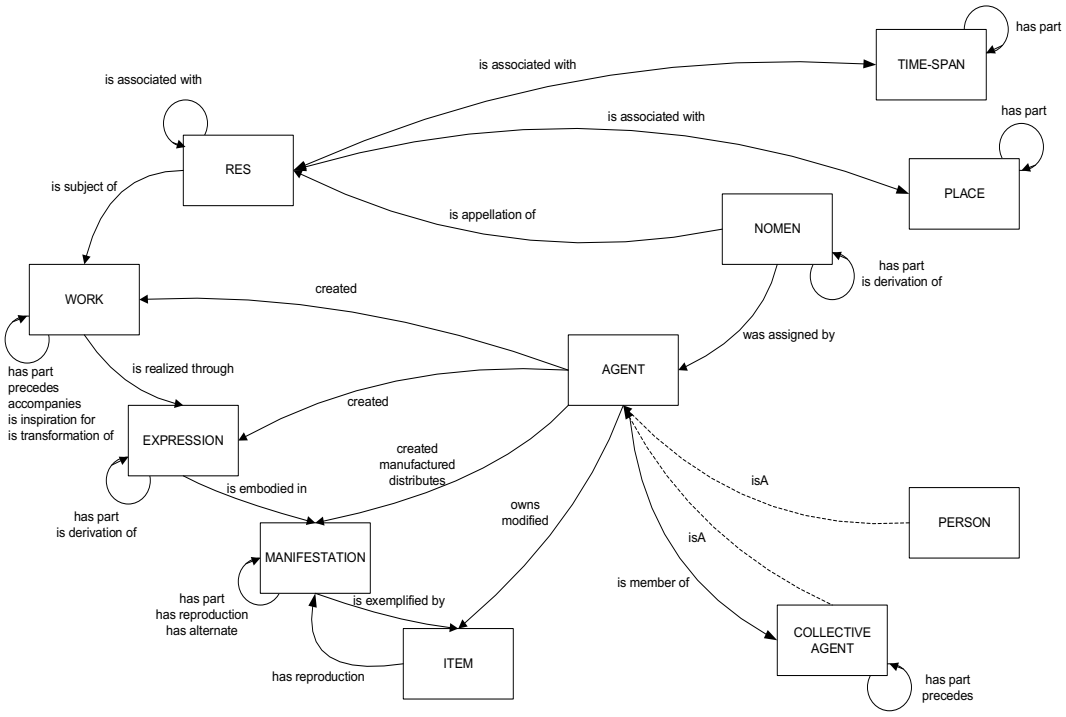


Figure 4: Overview of LRM relationships

4. Current status and future developments

In February 2016, the first stable draft of the LRM model was issued for a two-month world-wide review, according to IFLA practice. Subsequently the CEG incorporated revisions into the draft, which was then discussed by the full FRBR Review Group at its annual meeting in August 2016. The Review Group made decisions on all outstanding issues, leading to a final draft accepted at the FRBR Review Group level by the end of 2016. In accordance with the IFLA standards process, this final draft was submitted for approval to the IFLA Committee on Standards in April 2017 and posted on the IFLA website. The model was formally approved as an IFLA standard by the IFLA Professional Committee at its meeting held on August 18, 2017, prior to the IFLA World Library and Information Conference in Wrocław, Poland.

Several complementary documents have been issued, including a summary of changes in the model definition since the world-wide review draft. A *Transition Mappings* document, detailing the LRM equivalents for all user tasks, entities, attributes and relationships from the three previous models, is offered to guide the transition of any applications.

A working group of the ISBD Review Group has prepared a correspondence from the ISBD element set to LRM, discussed at its meetings in 2017, which is intended to lay groundwork for future revision of ISBD. Further mappings between LRM and other content standards are expected.

FRBRoo ver.2.4 (approved by IFLA in 2016) uses an object-oriented formalism to express the three FRBR Family models. The first steps towards bringing the object-oriented model into conformity with LRM took place in April 2017, at the Joint Meeting of the CIDOC CRM Special Interest Group and FRBR/CRM Harmonisation Working Group. The review, while not changing the nature of the model, will surely permit some simplifications and possibly lead to a "core" model for implementation. This work is ongoing, with a projected completion by the end of 2018.

As a general high-level model, LRM is intended to be expanded for implementation. LRM has already been adopted to guide the revision of *Resource Description and Access* (RDA), as part of the RDA Toolkit Redesign and Restructure (3R) project, which will demonstrate the methodology for extending the model.

References

Definition of FRBRoo : a conceptual model for bibliographic information in object-oriented formalism / International Working Group on FRBR and CIDOC CRM Harmonisation ; editors: Chryssoula Bekiari, Martin Doerr, Patrick Le Bœuf, Pat Riva. Version 2.4. November 2015. Available at: http://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf (accessed 2017-04-16) and as FRBR : object-oriented definition and mapping from $FRBR_{ER}$, FRAD and FRSAD, at: http://www.cidoc-crm.org/frbroo/sites/default/files/FRBRoo_V2.4.pdf (accessed 2017-04-16)

Functional requirements for authority data : a conceptual model / edited by Glenn E. Patton, IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). München : K.G. Saur, 2009. (IFLA series on bibliographic control ; vol. 34). As amended and corrected through July 2013. Available at: http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf (accessed 2017-04-16)

Functional requirements for bibliographic records : final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. München : K.G. Saur, 1998. (UBCIM publications ; new series, vol. 19). As amended and corrected through February 2009. Available at: http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf (accessed 2017-04-16)

Functional requirements for subject authority data (FRSAD) : a conceptual model / edited by Marcia Lei Zeng, Maja Žumer and Athena Salaba. München : De Gruyter Saur, 2011. (IFLA series on bibliographic control ; vol. 43). Available at: http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf (accessed 2017-04-16). Errata for section 5.4.2, October 2011, available at: http://www.ifla.org/files/assets/cataloguing/frsad/FRSADerrata2011.pdf (accessed 2017-04-16)

Guidelines for authority records and references. 2nd edition / revised by the IFLA Working Group on GARE Revision. München : K.G. Saur, 2001. (UBCIM publications ; new series, vol. 23). ISBN 978-3-598-11504-2. Available at: https://www.ifla.org/files/assets/hq/publications/series/23.pdf (accessed 2017-06-01)

Guidelines for subject authority and reference entries / Working Group on "Guidelines for Subject Authority Files" of the Section on Classification and Indexing of the IFLA Division of Bibliographic Control. München : K.G. Saur, 1993. (IFLA series on bibliographic control ; vol. 12). ISBN 10: 3-598-11180-0.

IFLA Library Reference Model : a conceptual model for bibliographic information / Pat Riva, Patrick Le Bœuf, Maja Žumer. August 2017. Revised after world-wide review, endorsed by the IFLA Professional Committee. Available at: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf (accessed 2017-09-01)

ISBD : International standard bibliographic description / recommended by the ISBD Review Group ; approved by the Standing Committee of the IFLA Cataloguing Section. Consolidated edition. Berlin ; München : De Gruyter Saur, 2011. (IFLA series on bibliographic control ; vol. 44). ISBN 978-3-11-026379-4.

Pisanski, Jan, Žumer, Maja. Mental models of the bibliographic universe : part 1 : mental models of description. J. Doc., 2010, vol. 66, no. 5, p. 643-667.

Pisanski, Jan, Žumer, Maja. Mental models of the bibliographic universe : part 2 : comparison task and conclusions. J. Doc., 2010, vol. 66, no. 5, p. 668-680.

RDA : resource description and access. ALA Publishing, 2009- . Available at: http://access.rdatoolkit.org/ (accessed 2017-09-11)

Riva, Pat, Žumer, Maja. The IFLA Library Reference Model, a step toward the Semantic Web. 83rd IFLA World Library and Information Congress, Wrocław, Poland, 2017. Available at: http://library.ifla.org/1763/1/078-riva-en.pdf (accessed 2017-09-11)

Transition mappings : user tasks, entities, attributes, and relationships in FRBR, FRAD, and FRSAD mapped to their equivalents in the IFLA Library Reference Model / Pat Riva, Patrick Le Bœuf and Maja Žumer. 2017. Available at: https://www.ifla.org/publications/node/11412 (accessed 2017-08-01)

UNIMARC manual – bibliographic format / edited by Alan Hopkinson. 3rd edition. München : K.G. Saur, 2008. (IFLA series on bibliographic control ; vol. 36). ISBN 978-3-598-24284-7.

The Use of Digital Object Identifiers in the National Diet Library Digital Collections

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; DOI; digital collections

Abstract

The National Diet Library (NDL) is the sole national library of Japan, and is responsible for the acquisition and preservation of as well as the provision of access to publications that comprise Japan's cultural and intellectual heritage. The NDL creates digital surrogates of its collections and acquires digital objects, to both of which it assigns Digital Object Identifiers (DOIs) as a means for ensuring permanent accessibility.

The NDL also serves on the board of the Japan Link Centre (JaLC), which is the only organization in Japan authorized as a Registration Agency (RA) for DOI. Since 2012, the JaLC has been managing the bibliographic and location information of academic resources held by Japanese institutions in electronic format, to which it assigns DOIs as a means of enhancing utilization and access from both inside and outside Japan. The NDL works together with the JaLC to promote widespread use of DOIs.

The NDL is proactively engaged in digitizing its holdings as a means of ensuring both preservation of and access to its collections. Both digital surrogates and born-digital materials can be searched and browsed via our digital archiving system, which is known as the National Diet Library Digital Collections. Assignment of DOIs to the digital objects stored in this system began in March 2014, and as of March 2017, more than 240,000 digital objects have been assigned DOIs in one of four categories:

1. Japanese doctoral dissertations that were received by the NDL between 1991 and 2000 and have been digitized
2. Books and periodicals published and digitized by the NDL
3. E-books and e-journals published by the NDL
4. Rare books and old materials digitized by the NDL

The assignment of DOIs to digital surrogates of rare books and old materials is a notable feature of the NDL's approach.

This presentation explains how DOIs are assigned at the NDL, the uses to which these DOIs are being put, and indicates the issues of DOIs assigned by the NDL.

References

Assignment of a DOI name to an object by NDL. Retrieved May 29, 2017, from http://www.ndl.go.jp/en/aboutus/doi.html.

Japan Link Center. Retrieved May 29, 2017, from https://japanlinkcenter.org/top/english.html.

National Diet Library Digital Collection. Retrieved May 29, 2017, from http://dl.ndl.go.jp/?__lang=en.

Data and Metadata Instantiation: Use Cases and a Conceptual Model

Presentation

Richard P. Smiraglia
University of Wisconsin-Milwaukee, USA
smiragli@uwm.edu

Keywords: data instantiation; metadata instantiation; DANS; clustering; disambiguation of groups; information retrieval; digital repositories

Abstract

Instantiation describes the phenomenon of variation in representation of information objects over time. Smiraglia (2008) describes it as the diatonic problem of both clustering and disambiguation of groups of what appear to be, but are not quite, iterations of the same object. Although the problem is well-known in bibliographic information retrieval (Smiraglia 2001), it also is well-documented among other kinds of information objects. Greenberg (2009) demonstrated instantiation among metadata records of evolutionary biologists, Coleman (2002) drew an analogy to instantiation among scientific models, and Smiraglia (2005) found instantiation among archival records of artifacts in a museum of archeology. As Greenberg points out (399), the problem is particularly acute in digital repositories where "automatic propagation, metadata inheritance, and value system adoption" contribute to a "lifecycle" that creates potentially ambiguous clusters.

Digital repositories are particularly susceptible to the problem of uncontrolled data and metadata instantiation because of the complex lifecycles of data deposit, use, and reuse. In repositories that require deposit of research data on a large scale, instantiation can become particularly acute. DANS (Data Archiving and Networked Services), a division of the Royal Netherlands Academy of the Arts and Sciences, is the "institute for permanent access to digital research resources" in The Netherlands (DANS 2017). The role of DANS is to encourage scholars to make their data accessible, interoperable and resusable, in a sustainable environment. In addition to serving as a host repository for tens of thousands of datasets, DANS also manages the NARCIS gateway to more than 160,000 datasets generated by Dutch scholars.

Recent research (Smiraglia and Park 2016) demonstrated one approach to a conceptual model of instantiation among open government data records, deriving core attributes "information object," "expression," "manifestation product type," "actor, "expression creation," and "information carrier" from the FRBRoo ontology of bibliographic instantiation. The proposed presentation combines these and other FRBRoo attributes with the generations of lifecycle modeling identified by Greenberg, as applied to a series of use cases from DANS.

Works Cited

Data Archiving and Networked Services (DANS). https://dans.knaw.nl/en/about accessed 12 June 2017.

FRBRoo (Functional Requirements for Bibliographic Records—object oriented), an extension of the CIDOC-CRM. http://old.cidoc-crm.org/frbr_inro.html# accessed 12 June 2017.

Coleman, Anita S. 2002. Scientific models as works. Cataloging & classification quarterly 33 no. 3/4:129-59

Greenberg, Jane. 2009. Theoretical considerations of lifecycle modeling: An analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. Cataloging & classification quarterly 47, no.3:380-402.

Smiraglia, Richard P. 2001. The nature of a work: Implications for the organization of knowledge. Lanham, Md.: Scarecrow Press.

Smiraglia, Richard P. 2005. Content metadata: An analysis of Etruscan Artifacts in a museum of archeology. Cataloging & classification quarterly 40 no. 3/4:135-51.

Smiraglia, RIchard P. 2008. A meta-analysis of instantiation as a phenomenon of information objects. Culture del testo e del document 9, no. 25:5-25.

Smiraglia, Richard P. and Hyoungjoo Park. 2016. Using Korean Open Government Data for Data Curation and Data Integration. Presentation, International Conference on Dublin Core and Metadata Applications, DC-2016, Copenhagen, Denmark. http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/447/509 Accessed 12 June 2017.

Linked Data I:
Transitions from Legacy

Using the Semantic Web to Improve Knowledge of Translations
Presentation

Karen Smith-Yoshimura
OCLC, USA
smithyok@oclc.org

Keywords: metadata; linked data; translations; multilingualism

Abstract

More than half of the almost 400 million bibliographic records in WorldCat are for languages other than English. Most of the monographs described were published only once. But a few million represent the core of our shared culture—works that have been translated into multiple languages, and sometimes translated multiple times into the same language. We learn about other cultures, and other cultures learn about ours, through these translations. As the world's largest bibliographic database, WorldCat is positioned to provide the translation history of works, using the W3C bib extension translationOfWork to communicate the relationship of each translation to the original work.

In our multilingual data enhancements project, our goal was to improve the descriptions of the most frequently published works, as they are the ones most likely to be translated and searched by users. In a database of MARC records, machine processes cannot support browsing or searching of works and their translations. Critical entities such as the title of the original work and the names of the translators are not always expressed in a machine-understandable form—and sometimes the information is missing altogether. Since a manual cleanup is not scalable, we explored the possibility of enriching MARC records with Linked Data from a third-party source, Wikidata. By integrating information from both WorldCat and Wikidata, we may be in a better position to present information about frequently-translated works in the preferred language and script of the user.

MARC records include data elements that can explicitly state that the record represents a translation, the language of the original and any intermediate translations, the title of the original work, and the translator(s) responsible for the translation. As long as *some* records accurately record this information, we can assert the correct relationships for the records that lack the information. Unfortunately, only a subset of all the relevant translations in WorldCat include such rich information. Many books written in non-Latin characters (such as Cyrillic, Greek, Russian) are often represented in WorldCat by the romanization only. This also makes the search by native-speakers unnatural and difficult. Using WorldCat records alone could not identify all the translations and their translators.

We enhanced the data retrieved from WorldCat with data retrieved from Wikidata by retrieving the Wikidata entries for a few works and its labels in multiple languages, even those written in non-Latin scripts. With the title/author match in a different language other than the original one, we can infer with high confidence a translation of the original work, even if the MARC record does not indicate it is a translation. The Wikidata entry often includes the non-Latin script for languages represented in WorldCat only in romanization. For example, we could use the label *Είναι και Χρόνος* from Wikidata rather than *Einei kai chronos* from WorldCat for the Greek translation of Heidegger's *Sein und Zeit.*

Data enrichment could be mutual. For example, Wikidata entries focus on the original title and do not describe all the translations represented in WorldCat; few Wikidata entries include translators, crucial to differentiate translations in the same language. Leveraging the strengths of each resource through linked data offers us the ability to present users an enriched view of our shared culture through translations.

Enhancing Metadata though Standardization and Validation: Practical Application at the University of Kansas Libraries

Presentation

Erin Wolfe
University of Kansas
Libraries, USA
edw@ku.edu

Keywords: Islandora; metadata enhancement; FAST Linked Data

Abstract

The Digital Initiatives department at the University of Kansas Libraries is in the process of migrating digital collections and assets to a locally hosted instance of Islandora to serve as our primary digital repository. As a key part of this process, we are taking the opportunity to clean up, standardize, enhance through linked data, and validate our metadata records prior to ingest in this new system.

Using a variety of open tools, we have developed a systematic and replicable method to create uniform metadata records that conform to our in-house guidelines and requirements. A final MODS record will serve as a master record for each object and is mapped to other schemas as appropriate, such as Dublin Core for display and OAI harvesting in Islandora.

Starting with metadata from a variety of sources, including MARCXML, ArchivesSpace EAD, and LUNA Imaging, XSL stylesheets transform the existing data to full MODS records. Then, a combination of Python scripts and OpenRefine's reconciliation service are used to convert LCSH terms to FAST Linked Data subject terms with URI attributes. Individual scripts are employed to create or update additional elements or values, such as Linked Data attributes for non-subject elements, multiple identifiers (e.g., institutional, Handles, ORCIDs, etc.), Creative Commons license statements, and other similar types of content.

In order to enforce compliance and consistency, we have developed a workflow that uses Schematron to (a) validate the record against the MODS schema, and (b) compare the contents of the record against different levels of required and preferred elements, as defined by the Digital Initiative's metadata guidelines, providing opportunity for continued improvements.

This presentation will highlight some of the processes that we are using and some challenges that we've faced. It will present a case study in practical application of linked data and systematized approach to metadata management in an academic library.

Extending Legacy Metadata with Linked Open Data

Jacob Jett
University of Illinois at Urbana-Champaign, USA
jjett2@illinois.edu

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

Alex Kinnaman
University of Illinois at Urbana-Champaign, USA
kinnama2@illinois.edu

Deren Kudeki
University of Illinois at Urbana-Champaign, USA
dkudeki@illinois.edu

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

Caroline Szylowicz
University of Illinois at Urbana-Champaign, USA
szylowic@illinois.edu

Abstract

Library special collections are valued by scholars and relied on to support both research and teaching. In recent years, libraries have invested heavily in digitizing many of these collections. Unfortunately, less effort and fewer resources have been expended post-digitization and many digitized library special collections today exist on the Web only in isolated information silos, difficult to find and disconnected from other resources that could provide users with valuable context. This begs the question: Can Linked Open Data (LOD) approaches be leveraged to help contextualize & enrich item-level descriptions of such collections and provide links to related information resources? This project report describes preliminary results from *Exploring the Benefits for Users of LOD for Digitized Special Collections*, a project still in progress which is examining this and related questions. Among the findings reported here: while special collections metadata are typically rich and ripe with LOD potential, the idiosyncratic nature of the collections and the metadata schemes used pose unique mapping and transformation challenges; the opportunity for adding links to item-level metadata is great, but finding links still requires significant cataloger involvement; at the scale of most library special collections, information from LOD sources can be retrieved in real time to enhance the presentation of items to end-users, providing context and links to related information. These findings suggest that the transformation of metadata into LOD and the inclusion in item descriptions of links can improve the connectedness of digitized special collections and enhance user interactions with these resources.

Keywords: metadata enhancement; linked open data; metadata mapping; schema.org

1. Introduction

After more than twenty years of digitizing special collections materials, libraries and other cultural heritage institutions have amassed a wealth of unique digital objects that span a broad spectrum of disciplines. As surrogates for primary sources and the products of scholarly research, the items in digitized special collections have significant value for further scholarship and pedagogy. However, studies suggest that, "Libraries are spending far more to create new [digital] resources than they are on maintaining and enhancing the ones they have already created" (Maron and Pickle 2013, p. 2). While the products of special collection digitization are often freely available on the Web, this lack of post-digitization investment and an overreliance on legacy descriptive metadata schemas and single collection finding aids designed originally for describing print resources has relegated many digitized special collections to information silos largely disconnected from the broader Web. This makes such resources difficult to discover and isolates them from potentially useful context. The need in the abstract for a more connected, shared cultural

heritage information space and to develop new methods of collection management and curation to overcome such 'siloization' tendencies has been recognized for several years, cf. the coordinated special issues of *Museum Management and Curation*, *Archival Science*, and *Library Quarterly* addressing this and related topics published at the end of the last decade (Marty 2008). Clearly there is a need to better connect digitized special collections to the broader Web and move away from information silos, but practical approaches to do this are still emerging and still being tested and evaluated. The outstanding question with regards to digitized special collections resources is what can be done after digitization to make these resources more discoverable by and more useful to users? Linking them to additional, related resources through Linked Open Data (LOD) (Burners-Lee 2006, Auer et al. 2007) is one potential value-added step that could be taken. However, our understanding of how LOD benefits digital libraries and their users is still limited (Corbet 2016).

Funded by the Andrew W. Mellon Foundation, the *Exploring the Benefits for Users of LOD for Digitized Special Collections*[1] project at the University of Illinois at Urbana-Champaign has been investigating LOD's potential to benefit both the stewards and users of digitized special collections. Two digitized theater collections (The Motley Collection of Theatre and Costume Design[2] and the Portraits of Actors, 1720-1920[3]) and a collection of digitized researcher notes encoded using TEI (the Kolb-Proust Archive for Research[4]) provide the foundation for the research project. This Project Report describes some of the unique challenges presented by special collections metadata and illustrates ways that LOD resources can be used to expand information presented to users and connect users to additional context external to the collection in order to make the resources contained in digitized special collections more useful for research and instruction.

2. Idiosyncratic data

One barrier to optimizing the utility of digitized special collections is the often imprecise and idiosyncratic nature of metadata records describing a collection's digitized objects or intellectual content. For instance, though the metadata schemes developed for the Motley and the Portraits of Actors collections were both based on simple Dublin Core, several extensions, re-interpretations of properties and classes, and other local scheme modifications were made when the objects were digitized and metadata initially transformed. Jett et al. (2016, 2017) found that besides conflating the objects being digitized with their digital representations (Woodley et al. 2005, Park & Childress 2009, Urban 2012), special collections metadata records often conflate descriptions of multiple individual entities and entities of different classes.

An import first-step towards ameliorating these problems is to choose a good linked-data-compliant vocabulary. At the time the project began Bibframe 2.0 was still under development and despite the existence of FRBR-oriented linked-data vocabularies—specifically $FRBR_{OO}$[5] and the SPAR family of ontologies[6]--schema.org[7] was selected for the project's mapping vocabulary of choice. The reason for this selection was entirely pragmatic: at the time the University library was experimenting with workflows for transforming much-richer MARC records into schema.org compliant metadata records, as was OCLC through their WorldCat Linked Data Vocabulary.[8]

2.1 Mapping Motley Image Metadata

For instance, in the Motley Collection exemplar case (Jett et al. 2016, p 3), metadata records describing costume sketches for a particular production of a play (e.g., as shown in Figure 1) were

[1] http://publish.illinois.edu/linkedspcollections/
[2] http://imagesearch-test1.library.illinois.edu/cdm/landingpage/collection/motley-new
[3] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/actors
[4] http://www.library.illinois.edu/kolbp/
[5] https://www.ifla.org/free-tags/object-oriented-frbr
[6] http://www.sparontologies.net/ontologies
[7] http://schema.org
[8] https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html

found to also contain metadata that described: a specific performance of the production (usually its opening night performance), the performance venue (theater), and the play itself (as a textual object). Persons associated with the production (e.g., producer, director, actor) were concatenated in a single field, individually identifiable by name and role only through punctuation and string-embedded labeling. Persons associated with the play as a written work (e.g., author, composer) were similarly concatenated in another metadata field.

Carefully mapping the Motley collection's legacy metadata model into the schema.org vocabulary (and any other linked-data-compliant vocabulary) begins by identifying these 1-to-1 violations in the existing metadata records. The resulting metadata records can then be divided into multi-part accounts that provide information particular to the different individual entities. Distinguishing the entities described is an essential first step in mapping legacy metadata into schema.org semantics and useful LOD-compatible descriptions. However, simply isolating and mapping the various entities described in the legacy metadata is not sufficient to make the metadata account optimal when displaying descriptions to users of these collections. The snippets of legacy metadata describing these additional entities must be enriched through the addition of links to external LOD services and other external resources (as described in Section 3 below).

FIG 1. Digitized costume sketch from Motley Collection with original metadata.

Since both the Motley collection and the Portraits of Actors collection are fairly traditional image special collections, the initial mapping of metadata records into Schema.org's vocabulary was relatively straightforward and should be adaptable for similar image collections. The primary class chosen to represent the collection's objects was that of schema:VisualArtwork. These objects (i.e., each sketch or photograph) were then disambiguated from the digitized image that represents them in online environments through use of the schema:associatedMedia relationship; in this way the schema:MediaObject (typically a schema:ImageObject) is distinguished in the LOD account

from the original hard copy Artwork (sketch or photograph). The play, production, venue and person entities are then isolated and given their own accounts in the LOD description through the use of the schema:isPartOf, schema:exampleOfWork, schema:locationCreated, schema:creator, and schema:contributor, relationships. This allows the LOD account to provide a series of true assertions, e.g., that the Visual Artwork is part of the Stage Work (the play's production), which in turn, is an example of a Book (the play) or Creative Work (when the play has not been formally published). A full account of the current mapping appears in Jett et al. (2016), though as the project is still ongoing, some further refinement of the mapping is anticipated.

2.2 Mapping Kolb-Proust Archive Data

In contrast to the two theater-focused collections, the digital objects in the Kolb-Proust Archive for Research (KPA) required a different approach. The KPA contains TEI-encoded transcriptions of two of Proust scholar Philip Kolb's research collections—*Bibliographie* and *Chronologie*. Each digital transcript provides the text that appears on a corresponding hardcopy notecard in these collections (see Figure 2). Each notecard in turn represents a single cohesive note made by Kolb in regard to an event or publication mentioned or alluded to in Proust's correspondence or of interest in regards to Proust's literary career or personal history. As the original editor of Proust's correspondence (spanning 21 volumes published between 1970 and 1993), Kolb's notes enabled him to sequence Proust's correspondence (most of which was not dated) and recognize the identity of people and events mentioned in his letters. These notecards provide potentially useful linkages between people, events and contemporaneous accounts of their interactions (e.g., in newspapers and other publications). In this sense, the cards are a sort of precursor to linked data. However, at first glance these transcriptions appear less well-aligned with the Schema.org ontology. Additionally, to express a wide range of complex relationships, succinct natural language was used on these cards; the transcription into TEI makes explicit (in a computational sense) only some of these relationships. To undertake a mapping of these archival collections, it was necessary to begin by carefully considering exactly what the transcriptions represent and contain vis-à-vis the classes of Schema.org. Person entities are prominently represented. Many notecards contain citations, which lend themselves to traditional metadata descriptive practices. However, capturing these names and citation in isolation loses the valuable context of the card on which they appear.

FIG. 2. A card from the KPA and its TEI transcription.

In this case the recognition of how to proceed finally arrived when we realized that each of Prof. Kolb's transcribed notecards could be represented using the schema:DataSet class which is a sub-class of schema:CreativeWork. This approach allowed us to accomplish three important things:

1. Preserve the overall context provided by each transcribed card (as just explained);
2. Use schema:citation to relate bibliographic entities to the card on which each appears;
3. Use schema:mentions to relate names and titles appearing elsewhere on cards.

Bibliographic works cited on cards were mapped to schema:CreativeWork or a subclass. We tried to use the appropriate level of granularity (i.e., schema:Article, schema:PublicationIssue, schema:Book, schema:PublicationVolume) whenever possible. But as in our earlier mappings (Jett et al. 2016, Han et al. 2017), it became clear that Schema could benefit from extensions that name other fine-grained classes for such bibliographic entities as "essays", "poems", and "short stories." In addition, it became clear that our earlier proposed extension for Schema, StageWork (Jett et al. 2016, p 3) would clearly benefit from its own set of finer-grained sub-classes that can better distinguish when a "Stage Work" is a "Play" versus when it is a "Variety Show" versus when it is an "Opera" and so forth.

3. Enriching Metadata with Links

A necessary step in order to migrate descriptions of items contained in digitized special collections from text-based metadata record formats to LOD-compatible serializations (of RDF graphs) is to identify and provide canonical URIs for entities described in the metadata being migrated. A common shorthand for this process is 'strings into URIs.' Note, a single entity, whether tangible (e.g., a person) or conceptual (e.g., a work) may be identified by multiple URIs. In the case of our project, the strings into URIs task has been completed for the Motley collection, and is still ongoing for the Portraits of Actors and KPA collections. For Motley the strings into URIs reconciliation process focused on people (author, composer, actor, director and other associated people), venues (theaters), subject headings (LCSH, TGMI and AAT), and the play / performance. As a precursor to searching for URIs some limited semi-automated metadata remediation was done, primarily de-conflating names and subject headings based on punctuation and formatting rules followed. As described below and once remediation was complete, we relied primarily on manual searching for links. Altogether 240 person-hours were spent on remediation and finding links for entities referenced in Motley metadata.

3.1 Named Entities (Persons) Identification & Reconciliation

URIs for persons mentioned in Motley metadata were sought using both automated and manual processes. This was done in order to ascertain which process worked better, to make an initial assessment of the labor required for the manual process and to help determine the degree to which the two approaches were complementary. Currently two of the most important LOD-compatible sources for person entity URIs are the Virtual International Authority File (VIAF)[9] and DBPedia[10]. Wikidata[11] is also becoming an important part of the Web's Linked Data landscape but at the time the project began we decided to focus our limited resources on VIAF which links to OCLC and other authority resources and DBpedia which is closely associated with Wikipedia. Both are related to other resources. DBPedia extracts structured content from the information found in Wikipedia[12]; Wikipedia URLs can be deduced from examination of DBPedia URIs and vice versa. VIAF draws on (and includes links to) many national catalogs and other library authorities, including the Library of Congress Name Authority File (LCNAF)[13], with which WorldCat Identities[14] also is closely synchronized. VIAF provides an interface for manually searching names, but we had better success (see Table 1) when manually searching using LCNAF. Manually searching for DBPedia URIs is best done by searching Wikipedia. Note that VIAF and Wikipedia have now been largely (but not

[9] http://viaf.org/
[10] https://dbpedia.org/
[11] https://www.wikidata.org/wiki/Wikidata:Main_Page
[12] https://www.wikipedia.org/
[13] http://id.loc.gov/authorities/names.html
[14] http://worldcat.org/identities/

fully) reconciled. Typically VIAF descriptions of persons include Wikipedia links when available and vice versa. For the names that we searched, we found that 10% of the Wikipedia and VIAF descriptions did not link to one another.

As shown in Table 1, manual searching for VIAF links yielded URIs for 218 of the 984 persons mentioned in the Motley metadata. Only 87 were found by searching VIAF directly, the rest were found by manually searching LCNAF and WorldCat Identities. By comparison, using the VIAF Auto Suggest API yielded URIs for 476 names, more than twice as many as found manually, but importantly, automatic searching in this way missed 106 URIs found by manual searching. Not surprisingly given that many individuals mentioned in Motley metadata were performers and directors rather than authors, we found Wikipedia / DBPedia URIs for 311 out of the same 984 persons; only some of which overlapped with VIAF URIs. We did limited experimentation with automating the searching of Motley names in DBPedia / Wikipedia, but in the absence of an API equivalent to the VIAF Auto Suggest API, we did not see much improvement over manual searching which was done first. For the Portraits of Actors Collection, automated searching, including of Wikipedia, was done first, which meant searching a smaller number of names manually. We expect to report quantitative results of this work soon. Our preliminary recommendation for similar collections is VIAF Auto Suggest API first, manual searching of LCNAF (limited to names not found using VIAF Auto Suggest), and lastly manual search of Wikipedia / DBPedia for names not found in the first two steps.

TABLE 1. Count of person URIs found through manual searching

Total persons identified in Motley metadata = 984 *Links have been found for 624 names*	**Count of URIs Found**
having Wikipedia / DBPedia links	311. (32%)
having VIAF links	218.* (22%)
found by searching viaf.org directly	87.
found by searching LC Name Authority File	196.
found by searching WorldCat Identities	93.
having Theatricalia links	475. (48%)
having IMDb links	353. (36%)
having IBDb links	42. (4%)
having more than 1 link	446. (45%)

* VIAF links for 476 persons (364 not found by manual search) were found using VIAF Auto Suggest

While both VIAF and DBPedia provide RDF descriptions of person entities, their scope is limited and their information about individuals is not complete. The scope of VIAF, for example, is biased toward authors of book-length publications. Since many of the individuals mentioned in the Motley metadata were actors, directors, producers, set designers, etc., we also manually searched several theater-related, non-LOD Web resources, specifically Theatricalia[15], IMDb[16] (the Internet Movie Database), and IBDb[17] (the Internet Broadway Database). While links (URLs) to these Web resources do not return RDF when de-referenced, they are a way for users to find more context about an individual. Because these resources focus on persons involved in theater and movie work, we found higher percentage of the names mentioned in Motley metadata in Theatricalia (58%) than in any other source, i.e., more than in either VIAF or Wikipedia. In addition to the ability to link with and acquire (albeit manually) metadata from these domain-pertinent resources provided additional, contextually important information. Theatricalia links allowed us, for example, to capture information regarding more specific roles that "Associated People" played with regards to a play's performance, production, or publication.

[15] https://theatricalia.com/
[16] http://www.imdb.com/
[17] https://www.ibdb.com/

3.2 Finding Links for Other Entities

VIAF and DBPedia / Wikipedia also proved useful sources for URIs for theaters (venue entities), while Theatricalia and DBPedia / Wikipedia were the best sources for URIs and URLs for performances and plays. Many theaters also have Web home pages that were easy to find through Google Search. Tables 2 and 3 give the quantitative results of manual searching to find URIs for theaters, performances and plays. The total number of entities involved was much smaller, i.e., although the Motley collection includes more than 4,700 set and costume design sketches, they are drawn from only 127 plays. Note that for this initial pass at adding URIs to the Motley metadata we conflated plays and performances. Based on feedback we may revisit this decision in the future.

TABLE 2. Count of theater URIs found through manual searching

Total theaters identified in Motley metadata = 59 *Links were found for 52 theaters*	**Count of URIs Found**
having Wikipedia / DBPedia links	49 (83%)
having VIAF links	45 (76%)
having home page links	36 (61%)
having other links	16 (27%)
having more than 1 link	47 (80%)

TABLE 3. Count of play / performance URIs found through manual searching

Total plays / performances identified in Motley metadata = 127 *Links were found for 105 plays / performances*	**Count of URIs Found**
having Wikipedia / DBPedia links	95 (75%)
having Theatricalia links	45 (35%)
having other links	10 (8%)
having more than 1 link	44 (35%)

4. Using LOD to Enhance Discovery and User Interaction with Collections

Transforming the structure of item-level metadata to better align with RDF and LOD best practices makes it easier for LOD-based applications to consume these resource descriptions; in theory, this should facilitate discovery. Similarly, including links to related Web resources and to LOD-based repositories containing additional information provides an opportunity to enhance user interaction with digitized special collections. While it is too early on this project to assess whether discovery of Motley resources (LOD-compatible metadata was only added to resource pages in late spring), work with the Google Structured Data Testing Tool (Figure 3) is encouraging.

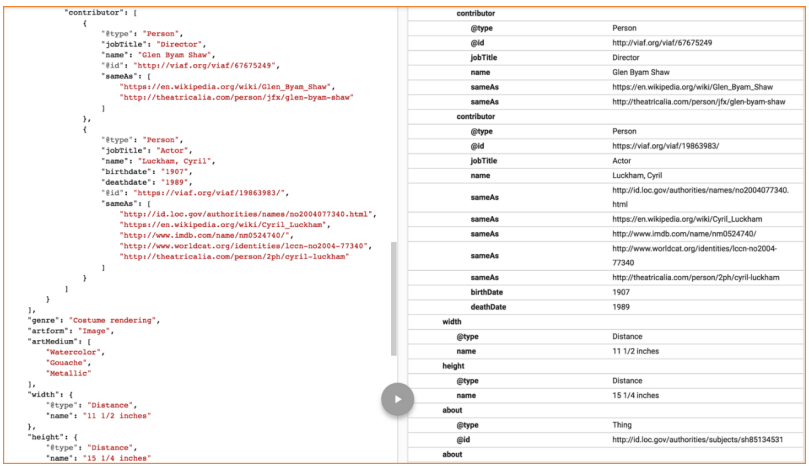


FIG. 3. Snippet of item-level RDF serialized in JSON-LD and viewed in Google Structured Data Testing Tool

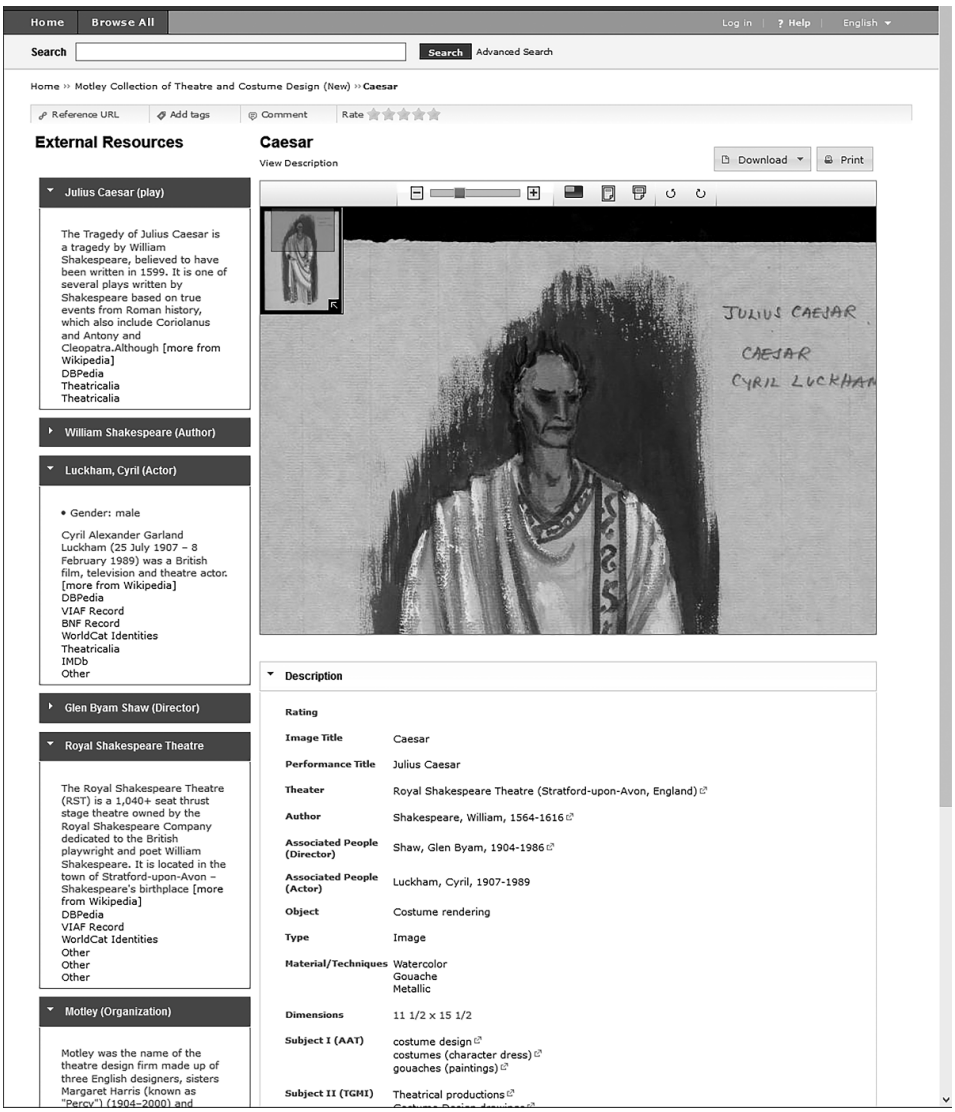


FIG. 4. Dynamically generated side-bar displaying links and context retrieved real-time from LOD services

To assess the viability and potential utility of the links added to Motley item metadata we took as inspiration the knowledge card approach increasingly being adopted by search engines like Google. Figure 4 showcases this approach as implemented. Here descriptions and links pertaining to the various entities mentioned in item-level RDF metadata (e.g., plays / productions, theaters, authors, directors, actors, etc.) are retrieved from LOD repositories (VIAF, DBpedia) and presented in boxes (1 per entity) along the left side of the browser window. The sidebar is produced using client-side JavaScript. The script, which runs on completion of the page load, parses RDF (serialized as JSON-LD and embedded in the Web page in a script element with a type attribute value of 'application/ld+json') to find URIs. These are then used to harvest links and selected metadata from LOD repositories (e.g., retrieving from DBpedia's page for the Royal Shakespeare Theatre[18] the value of the abstract[19]). URIs can be de-referenced directly if the LOD service implements the appropriate CORS (Cross-Origin Resource Sharing) headers or through proxies on our own server. Proxies offer the opportunity to implement short-term (e.g., 24 hour) caching to improve performance. The JavaScript also updates the default CONTENTdm metadata

[18] http://dbpedia.org/page/Royal_Shakespeare_Theatre
[19] http://dbpedia.org/ontology/:abstract

presentation, adding external links for entities described in the sidebar as well as for Getty[20] and Library of Congress[21] subjects.

The JavaScript to do this is relatively brief and uncomplicated, so to this point we have not needed to deploy a full-fledged Model-View-Controller (MVC) JavaScript framework (e.g., Angular.js, Backbone.js, Spine.js, to mention just a few), though our scripting does have JQuery.js dependencies. Nonetheless we have taken a model-view scripting approach in anticipation that we may wish to adopt a MVC framework library in the future. JavaScript object prototypes (illustrated in Figure 5) were defined for each class of entity in our RDF (person, theater, play / production) as our data model layer. Mustache.js templating was then used as our nascent view layer to manage the translation of these JSON data models into HTML.

```
function PersonEntity(id, identifier, homePage, longAbstract, shortAbstract, name, enWikiUrl, ViafUrl,
jobTitle, gender, birthDate, deathDate){
        this.id = id;
        this.identifier = identifier;
        this.homePage = homePage;
        this.longAbstract = longAbstract;
        this.shortAbstract = shortAbstract;
        this.name = name;
        this.enWikiUrl = enWikiUrl;
        this.ViafUrl = ViafUrl;
        this.links = new AutomapperConfig();
        this.type = 'Person';
        this.jobTitle = ( jobTitle ? jobTitle : 'Actor' );
        this.gender = gender;
        this.birthDate = birthDate;
        this.deathDate = deathDate;
    return;
    }
```

FIG. 5. The JavaScript prototype (data model) for person entities

A key benefit of the LOD-based architecture and implementation described above is that user interaction with item descriptions can be enriched without the need to redundantly store metadata values not required by the local system for search and discovery. This reduces risks of metadata staleness. A question remains as to what kinds of links and just-in-time contextual information at point of use would be most useful to users of the Motley collection. While we do not have scope within the current project to answer this question definitively, a small sample user test of the new interface illustrated in Figure 4 is currently underway. Results will be compared with those from a similar test conducted last year with the former user interface design illustrated in Figure 1. Note that user interface improvements between these tests was limited to changes stemming from the inclusion of LOD.

5. Conclusions & Future Work

This project report showcases one manner in which resources included in digitized special collections can be made potentially more useful through the addition of contextual information and links. As browser-based and mobile search engines become better at seamlessly integrating RDF-based, LO-compatible item descriptions LOD resources into their search indexes and search result displays through tabular facts, knowledge cards, and similar presentational arrangements, we can anticipate that the expectations of the users of traditional digital library systems will evolve. The identification of LOD resources through named entity reconciliation and subsequent harvesting of additional metadata and data for use by digitized special collections' users helps to meet some of

[20] http://vocab.getty.edu
[21] http://id.loc.gov

these evolving expectations while also better optimizing digitized collections materials for the more nuanced and data-rich environment that the Semantic Web is beginning to provide.

Linking out to general-purpose resources like Wikipedia or topically pertinent online repositories like the Internet Broadway DataBase (IBDb) or Theatricalia seem likely strategies for further optimizing digital library collections for use, though more in-depth and longitudinal assessments are required to confirm this. Such linking does carry with it risks that the linked to Web resource might go away at some point in the future. But these are typical risks for digital infrastructures and since no data is being fetched from these resources and rendered on-screen to the digital library's users, they are relatively minimal. Given the investment already made in digitizing the special collections with which we are working on this project, the incremental cost to enrich existing metadata with links and transform it into RDF seems reasonable; however, ultimately this determination will wait on further assessment of benefits. Finally in this project we are not attempting to assess in any systematic way the potential benefit that transforming legacy metadata into RDF may bring for machine, semantic-based reasoning. Obviously we have this in mind, especially in transforming the Kolb-Proust Archive for Research data from TEI to RDF using Schema.org semantics. Long term this remains a key motivation for exploring LOD for digital libraries. But for now, simply improving the connectedness of these collections will help us move away from the isolating information silos that have become all too ubiquitous.

Acknowledgements

We gratefully acknowledge the Andrew W. Mellon Foundation for generously funding the LOD for Digital Special Collections project. Though this research was supported in part by the Mellon Foundation, the opinions, conclusions and findings expressed are those of the authors and do not necessarily reflect the views of the Foundation.

References

Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web. Lecture Notes in Computer Science 4825*, p 722. doi:10.1007/978-3-540-76298-0_52.

Berners-Lee, T. (2006). Linked Data. Design Issues. [revised 2009]. Retrieved from: https://www.w3.org/DesignIssues/LinkedData.html.

Corbett, L. E. (2016). Linked data advice anyone? (Who uses Google?) *Technicalities 36*(1), pp 1-7.

Jett, J., Han, M.-J., and Cole, T. W. (2016). Disambiguating descriptions: Mapping digital special collections metadata to Linked Open Data formats. Poster presented at the *79th ASIS&T Annual Meeting* (14-18 October 2016, Copenhagen, Denmark).

Jett, J., Cole, T. W., Han, M.-J., and Szylowicz, C. (2017). Linked Open Data (LOD) for library special collections. Poster presented at *17th Annual ACM/IEEE-CS Joint Conference on Digital Libraries* (19-23 June 2017, Toronto, Canada).

Maron, Nancy L. and Pickle, Sarah. (2013). *Appraising our Digital Investment: Sustainability of Digitized Special Collections in ARL Libraries*, Association of Research Libraries and Ithaka S+R, 2013. Retrieved from http://sr.ithaka.org/research-publications/appraising-our-digital-investment.

Marty, P.F. (2008). An introduction to digital convergence: libraries, archives, and museums in the information age. *Archival Science* 8: 247-250. doi:10.1007/s10502-009-9094-1. Concurrently published: (2009) *Museum Management and Curatorship* 24 (4): 295-298, doi: 10.1080/09647770903314688; and (2010) *Library Quarterly* 80 (1): 1-5, doi: 10.1086/648549.

Park, J. and Childress, E. (2009). Dublin Core metadata semantics: An analysis of perspectives of information professionals. *Journal of Information Science 35*(6), pp 1-13.

Urban, R. J. (2012). Principle paradigms: Revisiting the Dublin Core 1:1 Principle. Doctoral Dissertation. Retrieved from: http://hdl.handle.net/2142/31109

Woodley, M. S., Clement, G., and Winn, P. (2005). DCMI Glossary. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/usageguide/glossary.shtml

Linked Data II:
In and Around the Library

Metadata for the Energy Performance Certificates of Buildings in Smart Cities

Ana Alice Baptista
Centro ALGORITMI, Universidade do Minho, Guimarães, Portugal
analice@dsi.uminho.pt

Sara Catarina Silva
Centro de Computação Gráfica, Guimarães, Portugal
a69185@alunos.uminho.pt

Abstract

SusCity is a MIT Portugal project that falls within the scope of smart cities. One of its tasks aims to research and develop metadata artefacts to be used in the scope of a Linked Open Data platform. In this article, we report the process and results associated with the development of the following metadata artefacts: an application profile, a metadata schema and four controlled vocabularies. The application field is the energy certification of buildings. For the development of the application profile, we inspired ourselves in the Me4MAP method although we did not use it thoroughly. The creation of the metadata schema and controlled vocabularies involved the use of Wikidata, so all new terms (RDFS classes and properties and SKOS concepts) are related to Wikidata terms. The results include the application profile, the metadata schema and the controlled vocabularies. The application profile has 13 properties, four of which are new. The controlled vocabulary on measures for energy performance has 22 new terms spread over four levels. The remaining controlled vocabularies just hold a few terms each. All the artefacts are open to the community for use and reuse.

Keywords: smart cities, energy performance certificate, metadata, linked open data, application profile.

1. Introduction

SusCity (Urban data-driven models for creative and resourceful urban transitions) is an MIT Portugal project in the scope of smart cities. The MIT Portugal program "is a strategic partnership between Portuguese Universities and Research Centres, the Massachusetts Institute of Technology and partners from industry and government (...) [and] its goal is to strengthen the country's knowledge base and international competitiveness through a strategic investment in people, knowledge and ideas in innovative technology sectors" ('MIT Portugal', 2017). The development of the concept of a "smart city" rises from a complex association between technology, society, economy, administration and politics. A smart city is a city that invests in human and social capital, traditional (transport) and modern (ICT) infrastructures to enable a sustainable economic development and ensure a high quality of life with the intelligent management of natural resources (Caragliu, Del Bo, & Nijkamp, 2011). SusCity "is focused on developing and integrating new tools and services to increase urban resource efficiency with minimum environmental impacts while contributing to promote economic development and preserving the actual levels of reliability" (SUSCITY, 2016).

The project is divided into 6 (six) work packages (WPs), from which WP1, WP3, WP4 and WP5 generate data that is fed into the data processing platform created in the scope of WP2 (see FIG. 1). The data processing platform has several constituents, two of which are the Analytics module and the Linked Open Data (LOD) platform. The internal analytics module provides mechanisms for analysing vast amounts of consolidated data. The LOD platform is where linked data is made openly available and for which we are currently developing several metadata application profiles (MAPs), controlled vocabularies and metadata schemas. This article reports

on the development process and outcomes of a MAP, one metadata schema and four controlled vocabularies related to energy performance certificates (EPCs) of buildings.

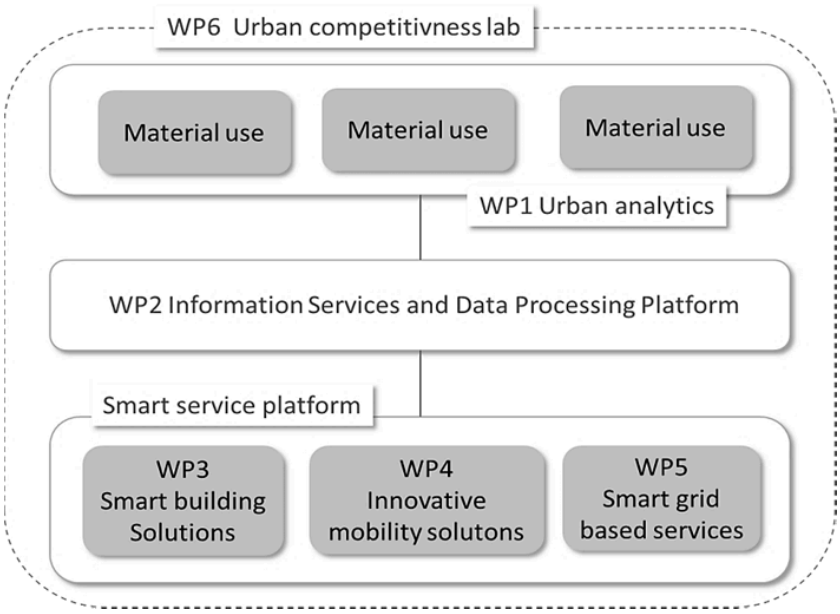


FIG. 1 - SusCity project structure (FCT Web site as cited by Aelenei et al., 2016).

In the past years, the world has seen a rapid growth in energy consumption. This growth has raised concerns related to exhaustion of energy resources and environmental impacts. Buildings are believed to be one of the main contributors to the increase in power consumption, exceeding major sectors such as industry and transportation (Pérez-Lombard, Ortiz, & Pout, 2008). Therefore, energy performance of buildings must be a top priority at a national and international level. Increase the energy performance of buildings can be achieved during the architectural design when choosing the building's mechanical system or with a smart management of the building utilities (Sozer, 2010). It can also be achieved by adopting maintenance measures specifically targeted to increase the energy performance of the building. Space heating is believed to be the most energy consuming end use (71%), followed by water heating (12%), cooking (4%) and lighting, air conditioning and other appliances (15%) (Odyssee-Mure, 2012).

In 2016, we and others (see acknowledgements) developed a research work on Linked Open Data, energy and energy efficiency of buildings. That work aimed at identifying, under these topics, 1) the main international and European standards; 2) metadata schemas and controlled vocabularies used in open datasets in the cities of Paris, London and Amsterdam; and 3) the metadata elements, encoding schemas and units of measure used in the datasets and catalogues identified in 2). The searches were made using the following services and websites: Google Scholar, Google, Linked Open Vocabularies (LOV), European Data Portal, and the Web open data portals of the cities mentioned above. As we were having difficulties to find the information we were looking for, in some cases we expanded the searches to other open data portals. Unsurprisingly, we were able to find legislation, international or European standards and guidelines regarding the energy efficiency of buildings and regarding Linked Open Data (LOD) separately, but almost nothing regarding the two combined. The little we found is limited to specific datasets encoded in RDF but none of them had the specific information we needed.

Because we could not find the artefacts we needed, we had to create them. We started by developing the MAP for the EPCs. The development of this MAP eventually required other developments: a metadata schema that would house new properties and classes, some of them

specific to this domain and three small controlled vocabularies to constrain the range of three properties. Yet, another development was made as a request from our partner, ADENE - Agência para a energia (agency for the energy): the encoding in SKOS (Simple Knowledge Organization System) of a controlled vocabulary about measures to improve the energy performance of buildings. These developments required some design choices: whenever possible, we have given priority to simplicity and potential interoperability to the detriment of the wealth of semantic description. The work reported here has been done in close collaboration with ADENE, which is responsible for the management and operation of the Buildings' Energy Certification System, which has already 10 years of existence and has issued more than 1,250,000 energy certificates (ADENE, 2017).

This article is divided into four sections, starting with this introduction followed by a section on the data and methodological procedures, where we characterise the data that we had available and the methods and techniques used for the development of the metadata artefacts. Then we present the results in three different subsections: subsection 3.1 for the constraints matrix of the MAP, subsection 3.2 for the metadata schema and subsection 3.3 for the controlled vocabularies. In section 4 we present the final remarks and future work. The prefixes and URIs of the namespaces are made available before the references.

2. Data and methodological procedures

2.1. The data

We are opening two kinds of data: data sent by project partners and data generated within WP2 by the internal analytics module. In most of the cases, data sent by partners cannot be used in full because it can lead to privacy or security issues. For example, we cannot open data on energy consumption of each dwelling because besides being itself private data, it could allow the inference of other private information concerning those buildings or even concerning their residents (e.g., periods of time when those citizens are at or out of home). The same goes for security concerns: certain information, such as the architectural plans of buildings, or the building materials, could be used for purposes that could jeopardise the safety of their residents (e.g., terrorist attacks). For these reasons, when developing the MAP, it was necessary to eliminate some details about the data that came to our hands. On the contrary, the data generated within the WP2 internal analytics module already arrived anonymized and ready to be open, so there was no need to foresee changes when designing the MAP.

At their website, ADENE makes openly available for human consumption a fraction of the data related to individual EPCs (see http://www.adene.pt/sce/micro/certificados-energeticos). The data related to this article is part of that fraction as ADENE considered that because machine readable data is processable and easily relatable to other data, information about individual buildings should not be disclosed in a machine readable way. Therefore, door and floor numbers of apartments have not been entered into the MAP. Because of this and because EPCs already have parts of the address as standalone fields (e.g., Parish or Municipality), we decided to use the attribute "street" instead of "address" which was the one originally present at the EPCs. ADENE also considered that the name and number of the energy expert that analysed the building should not be opened in a machine readable way. Once that the data about the EPCs is to be shared globally, we decided to add the attribute Country. Therefore, the data to be opened for each EPC was the following:

- Certificate number - number that identifies the EPC;
- Document type - EPCs can be of two types: Regulatory Compliance Statement or Energy Certificate;
- Date of issue - the start date of the EPC;
- Validity date - the last day for which the EPC is valid for;
- Energy label - label that represents the energy performance;
- Country - country of the building;
- District - district of the building;
- County[1] - county of the building;
- Civil parish[2] - parish of the building;
- Street - street of the building;
- Land registry office - office that registers the ownership of land and property;
- Land registry entry number - number that identifies the property in the land registry office;
- Type of use of building - The type of use given to the building.

An EPC should clearly portray the energy performance situation of the building and should provide recommendations to the owner to improve the energy performance of the building in question (European Parliament and the Council of the European Union, 2003). One of the controlled vocabularies that were encoded in SKOS represents the improvement measures that ADENE uses in their recommendations to owners. In our case, as we are only dealing with a small fraction of the EPC's data, this data does not have a direct relationship with the controlled vocabulary of improvement measures, as they do not directly link to each other. Nevertheless, ADENE requested that we encoded it in SKOS, so that it could be made available for use and reuse both by them and by other organisations. The data and the structure of the controlled vocabulary were provided by ADENE in a word file.

2.2. Methods and techniques

When we started the development of the MAP we already knew what data we would release, so we did not need to go through the vast majority of steps that others perform and that are foreseen in the Singapore Framework (Nilsson, Baker, & Johnston, 2008) or in Me4MAP (Malta & Baptista, 2013). For example, we did not need to define the functional requirements as they had already been defined for the human interfaces (Costa & Santos, 2015; 2016) - our mission was just to make widely available in a machine readable way the data that was being used in human interfaces. Similarly, we could have designed our constraints matrix without having designed the domain model. However, a precise specification of the domain model is relevant in cases like this, since it helps us to keep in mind the context of the properties we intend to use. It is also appropriate for documentation purposes. In our case, we started the design of the MAP in the domain model (see FIG. 2).

[1] The administrative division of the territory varies from country to country. The word município in Portuguese refers to a place with a certain degree of administrative autonomy and run by a local government. It aggregates several freguesias. Several municípios are part of a district. For the sake of interoperability we will use the word "County" although we know the meaning may not be exactly the same.

[2] In Portugal, the smallest administrative unit is called freguesia and it may correspond with what in some countries is called civil parish, term that we will use.

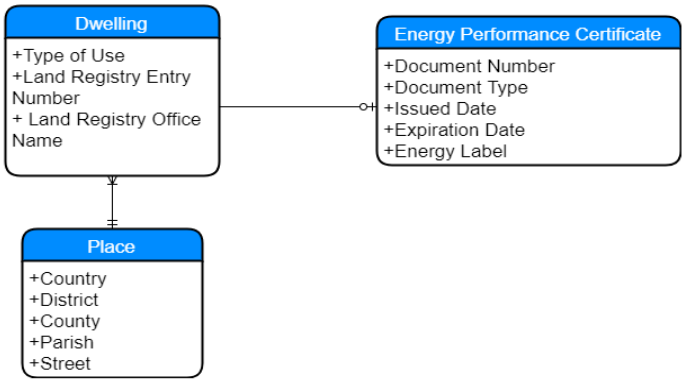


FIG. 2 - Domain model of the EPCs data opened in the scope of the SusCity project.

After designing the domain model, we looked for properties that matched its attributes. The majority of the searches for properties were done using LOV, which is a catalogue of the available vocabularies that allow the description of data on the Web (Vandenbussche, Atemezing, Poveda-Villalón, & Vatant, 2017). The identified schemas were also used to look for properties. One important aspect when choosing a property is to verify if it has restrictions (domains or ranges) that are incompatible with the intended use in our application. Although we are aware that this is not something taken into consideration by some developers, we think that it is an important aspect of the semantics of properties and we did not want to neglect it. Sometimes we have come to situations where we have had to do without something: either the semantic rigor of the definition of the property in the original schema, or the choice of a property of a widely used schema, or the respect for the rigor of the restrictions (compatibility between domain or the MAP range and the original domain or range of the property).

As an example, we have the attribute Freguesia (Civil Parish) for which we could have used the DBpedia property `parish`. However, this property has as domain the DBpedia class `PopulatedPlace` that is defined as "a place or area with clustered or scattered buildings and a permanent human population " (see http://dbpedia.org/ontology/PopulatedPlace). This domain makes all sense but not in our case, as the domain for this property should be left unspecified or it should be a document: the EPC. This apparent nonsense results from a design simplification: it is not an EPC that has a direct relationship with a given parish, but a building that is placed in a local that corresponds to that parish. We could have opted for having three classes, EPC, Dwelling and Place, which could be related by some properties such as Dublin Core `relation` or one or more of its sub-properties. If this were the case, the DBpedia property `parish` would apply well as its domain fits well with the class we could have defined (our Place would also be a populated place). However, we preferred to dispense with some semantic rigour in the description of resources and gain simplicity, mainly in the future processing of the resources' descriptions: it will not be necessary to execute complex queries to obtain all the required information as all the properties are constrained in the MAP to have as domain just one class (the Energy Performance Certificate), which implies that we have just one description per EPC. Summarizing, since we do not know the future users of our data, we endeavoured to opt for the simplicity of the descriptions and the possible semantic rigour in the use of the properties in what regards their constraints.

We had similar problems in what regards ranges of properties, but in some of these cases we dispensed with the respect for the original range, otherwise we would have a multiplication of properties and metadata schemas in our MAP. Our option, then, was to use just one property per attribute, even if we had to dispense with the rigour of ranges. For example, the `vcard:country-name` property has as range `xsd:string` and we are using it as a datatype property and also as an object property.

For each attribute that we wanted to use, we looked for a LOD property that corresponded to it. In cases where we found more than one candidate property that could be applied, we used the following priorities of choice: 1) property in a schema corresponding to some standard or some recommendation of some relevant entity in the field of application (energy certification) or in LOD; 2) property widely used by the metadata community; or, finally, 3) property defined in a formal schema and used by at least two datasets. Only in cases where we could not find properties that satisfied at least one of these criteria, we decided to create new properties. We then filled a table, the constraints matrix, with the following columns related to properties: attribute original name, label, namespace, property name, original domain, original range, domain in the MAP, range in the MAP, and cardinality (see Table 1 and Table 2). This constraints matrix is very different from those advocated by Coyle and Baker (2009) and by Malta and Baptista (2013). We have made the changes we deemed necessary to meet our needs for description and representation. Therefore, comparing do Malta and Baptista's table:

- We inserted a new column, "Attribute Original Name", that refers to the original name of the attribute in the domain model, so that the transition from the domain model to the constraint table is clear and evident.

- We kept the column "Label" (in English).

- The column "Property" was broken down in two columns: "Namespace" and "Property Name", so that we could grasp quickly which were the namespaces being used.

- We inserted two new columns, "Original Domain" and "Domain in the MAP", so that we could easily verify the compatibility of the properties' domains. The domain in the MAP should be a domain that is compatible with the original domain, i.e., it should be the same of the original domain or one of its sub-domains. For example, we can define as domain a specific type of document when the original domain is a general document.

- The column "Range" was broken down in two columns: "Original Range" and "Range in the MAP", so that we could easily verify the compatibility of the properties' ranges. The range in the MAP should be a range that is compatible with the original range, i.e., it should be the same of the original range or one of its sub-ranges (as you will see later, this was not always achieved). For example, we can define as range a controlled vocabulary in cases where the original range is unspecified.

- We removed the columns "Value String", "SES URI", "Value URI" and "VES URI" as their contents may be placed in the column "Range in the MAP".

- The columns "Min" and "Max" were replaced by the column "Cardinality", which means the number of times a given property may be used in a single description.

- We removed the columns "Type", "Usage" and "Related Description". We are using the "Range in the MAP" column to specify the type and we are specifying the usage at the level of the domain model. The "Related Description" is unnecessary in our case as we are only using one class. If we were using more, we could have inserted a new property for the relation and specify the related class in the "Range in the MAP" column.

The development of this MAP revealed the need to create a metadata schema with four new properties and a class. This schema was created using the Resource Description Framework Schema (RDFS). Equivalent terms for the properties and the class were created in Wikidata for two reasons: 1) to have them open for definition, evolution and control by the crowd (live properties); 2) expectation of having more stable URI's that assure long term existence of the terms; which result in 3) expectation of greater interoperability in the future. However, we created them with minimum structure so that they were not unnecessarily constrained. The structure was instead defined in the RDFS file.

For terms' names, Wikidata provides a character string mainly constituted of numbers, i.e., by reading the term name through the URL humans cannot immediately infer its semantics, even in a general way. Due to this fact, we created easily human readable term names in our RDFS file and

then related them to the Wikidata terms through `owl:equivalentClass` and `owl:equivalentProperty` properties (OWL - Web Ontology Language). We also did not define domain and range restrictions to these new properties, leaving them open to different usages. The restrictions were instead defined at the MAP level. This facilitates the reuse of terms across different schemas while safeguarding minimal semantics.

Similarly, regarding the controlled vocabularies, all SKOS concepts have equivalent terms in Wikidata. The SKOS and Wikidata terms were related via `owl:sameAs` property. We also chose to establish in Wikidata the minimum relations between the concepts and to represent most of the vocabulary structure in the SKOS files. This facilitates the reuse of terms across controlled vocabularies while safeguarding minimal semantics.

3. Results

3.1. Constraints matrix

Screenshots of the constraints matrix are presented in Table 1 and Table 2. The constraints matrix is fully available online at http://hdl.handle.net/1822/46455.

TABLE 1. Constraints matrix for the EPC SusCity Metadata Application Profile (part 1).

Attribute original name	Label	Namespace	Namespace prefix	Property name
Nº de documento (certificado)	Document number	http://dbpedia.org/ontology/	dbpedia-owl	documentNumber
Tipo de documento	Document type	http://purl.org/dc/terms/	dct	type
Data de emissão	Date of Issuance	http://purl.org/dc/terms/	dct	issued
Data de validade	Date of Validity	http://purl.org/dc/terms/	dct	valid
Classe energética	Energy Label	http://opendata.dsi.uminho.pt/terms/energy/EPCterms/	ec	energyLabel
País	Country	http://www.w3.org/2006/vcard/	vcard	country-name
Distrito	District	http://data.ordnancesurvey.co.uk/ontology/admingeo/	osadm	district
Concelho	County	http://data.ordnancesurvey.co.uk/ontology/admingeo/	osadm	county
Freguesia	Parish	http://data.ordnancesurvey.co.uk/ontology/admingeo/	osadm	parish
Rua	Street	http://www.w3.org/2006/vcard/	vcard	street-address
Conservatória do Registo Predial	Land Resgistry Office	http://opendata.dsi.uminho.pt/terms/energy/EPCterms/	ec	landRegistryOffice
Artigo matricial nº	Land Registry Number	http://opendata.dsi.uminho.pt/terms/energy/EPCterms/	ec	landRegistryNumber
Tipo de uso do edifício ou fracção	Type of use of building	http://opendata.dsi.uminho.pt/terms/energy/EPCterms/	ec	useOfBuilding

Some attributes of ADENE's EPCs (see http://www.adene.pt/sce/micro/certificados-energeticos) were not considered during the development of the MAP because of privacy and/or security reasons as reported above. These attributes are "address" and "building unit". We decided that the address would be left at the street level (buildings and dwellings are not identified). As we already had some components of the address as standalone attributes (Country, District, Municipality, Parish), we just needed to include the attribute Street instead. Also, as reported above, ADENE decided to leave out the name and number of the expert who issued the EPC. Some of the choices we had to make when creating this table are pretty straightforward and do not need further explanations.

TABLE 2. Constraints matrix for the EPC SusCity Metadata Application Profile (part 2).

Namespace prefix	Property name	Original domain	Original range	Domain in the MAP	Range in the MAP	Cardinality
dbpedia-owl	documentNumber	dbo:Document	xsd:string	ec:EnergyPerformanceCertificate	xsd:string	1
dct	type	Unspecified	rdfs:Class	ec:EnergyPerformanceCertificate	string, EPC Types vocabulary (http://opendata.dsi.uminho.pt/terms/energy/EPCTypes.skos)	1-2
dct	issued	Unspecified	rdfs:Literal	ec:EnergyPerformanceCertificate	xsd:date	1
dct	valid	Unspecified	rdfs:Literal	ec:EnergyPerformanceCertificate	xsd:date	1
ec	energyLabel	Unspecified	Unspecified	ec:EnergyPerformanceCertificate	xsd:string + Energy Labels vocabulary (http://opendata.dsi.uminho.pt/terms/energy/energyLabels.skos)	1
vcard	country-name	Unspecified	xsd:string	ec:EnergyPerformanceCertificate	xsd:string, TGN, GeoNames	1-3
osadm	district	Unspecified	osadm:District & others	ec:EnergyPerformanceCertificate	xsd:string, TGN, GeoNames	1-3
osadm	county	Unspecified	osadm:County & others	ec:EnergyPerformanceCertificate	xsd:string, TGN, GeoNames	1-3
osadm	parish	Unspecified	osadm:CivilParish & Others	ec:EnergyPerformanceCertificate	xsd:string, TGN, GeoNames	1-3
vcard	street-address	Unspecified	xsd:string	ec:EnergyPerformanceCertificate	xsd:string	1
ec	landRegistryOffice	Unspecified	Unspecified	ec:EnergyPerformanceCertificate	xsd:string, New controlled vocabulary to be encoded in SKOS	1-2
ec	landRegistryNumber	Unspecified	Unspecified	ec:EnergyPerformanceCertificate	xsd:string	1
ec	useOfBuilding	Unspecified	Unspecified	ec:EnergyPerformanceCertificate	xsd:string, EPC Types of Buildings vocabulary (http://opendata.dsi.uminho.pt/terms/energy/EPCTypesOfBuildings.skos)	1-2

However, the following cases need a clarification:

- The `dbpedia-owl:documentNumber` property for the Document Number attribute has as its original domain is `dbo:Document`. The new class `ec:EnergyPerformanceCertificate` is encoded in the metadata schema as subclass of (`rdfs:subClassOf`) `dbo:Document`. The domain of this property in our MAP is then restricted to `ec:EnergyPerformanceCertificate`. The original domain of the rest of the properties is unspecified, but in this MAP all of them are restricted to `ec:EnergyPerformanceCertificate`.
- The properties `dct:type`, `ec:energyLabel` and `ec:useOfBuilding` have as range the terms of controlled vocabularies created by us and encoded in SKOS (see next section).
- The OSADM schema does not provide properties for the Country and Street attributes, reason why we used another schema. VCard was chosen because it is widely known and used. Despite we tried, we could not find the five address related properties in a single metadata schema especially due to the peculiarity of attributes like County or Civil Parish.
- The properties `vcard:country-name`, `osadm:district`, `osadm:county` and `osadm:parish` will be used having as range in the MAP the terms of the Thesaurus of Geographic Names (TGN), GeoNames and/or a value string.
- The properties `ec:energyLabel`, `ec:landRegistryOffice`, `ec:landRegistryEntryNumber` and `ec:useOfBuilding` were created by us in RDFS.
- The property `ec:landRegistryOffice` has as range a value string and terms of a controlled vocabulary to be encoded in SKOS in the future.

3.2. Metadata schema

The new metadata schema has four new properties and a new class. The following properties have been created in RDFS:

- Land Registry Office (conservatória do registo predial), property: `landRegistryOffice` - office that registers the ownership of land and property;
- Land Registry Entry Number (Artigo matricial nº), property: `landRegistryEntryNumber` – number that identifies the property in the land registry office;
- Type of building or building part (Tipo de uso de edifício ou fracção), property: `UseOfLandOrProperty` – The type of use given to the LandOrProperty;
- Energy label (classe energética), property: `energyLabel` – label that represents the energy performance of some thing.

The following class has been created in RDFS:

- Energy Performance Certificate, class: `EnergyPerformanceCertificate`.

The domain and range of all properties were left unspecified. The new metadata schema is available at http://opendata.dsi.uminho.pt/terms/energy/ .

3.3. Controlled vocabularies

We have created and encoded four controlled vocabularies in SKOS. Three of them derive directly from the development of the MAP and have the following terms:

- For the type of document, the EPC Types vocabulary:
  - `Regulatory Compliance Statement`
  - `Energy Certificate`
- For the type of use of the building, the EPC Types of Buildings vocabulary:
  - `Domestic`
  - `Non-domestic (small commercial and service buildings and large small commercial and service buildings)`
- For the energy labels, the Energy Labels vocabulary:
  - Terms that coincide with the energy labels currently considered by ADENE, as shown in Figure 3.

FIG. 3 - Energy labels considered by ADENE (see http://www.adene.pt/sce/o-que-e-1).

The Energy Labels vocabulary currently only has the European energy labels. Other energy labels may be added in the future. It would be interesting if, at the MAP level, we could constrain the range of a property to a part of a controlled vocabulary. This would allow us to constrain the range of the `ec:energyLabel` property to the EU branch of the energy label vocabulary.

The controlled vocabulary on measures to increase the energy performance of buildings has 22 new terms spread over four levels. Details of the terms are still being worked out by our partner , ADENE. Figure 4 shows the tree structure of the vocabulary.

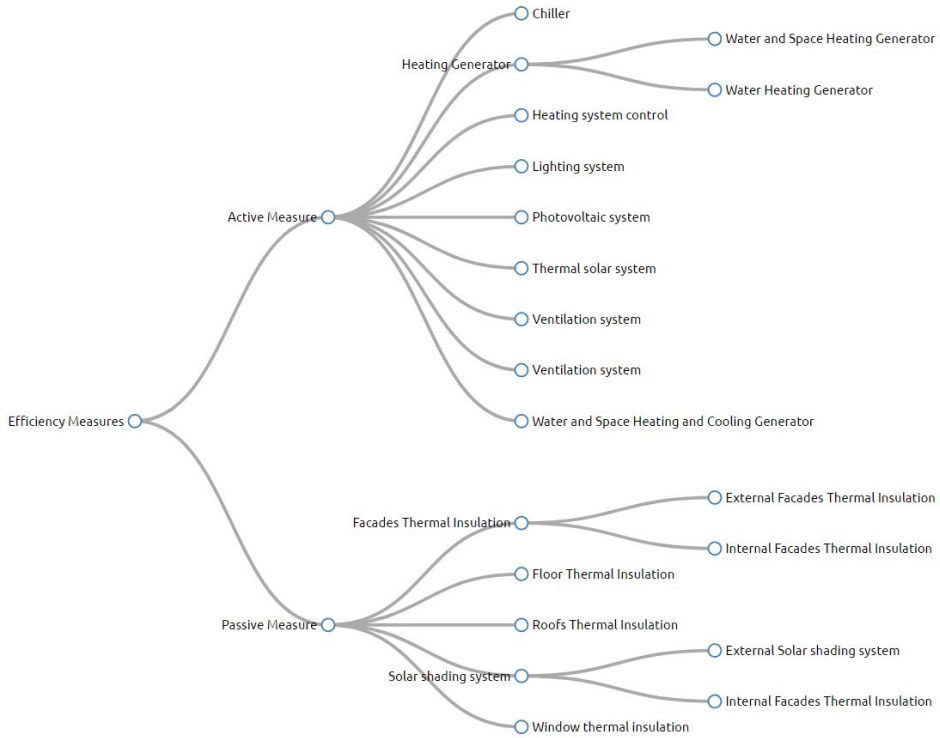


FIG. 4 - Tree structure of the vocabulary of measures for energy performance.

These controlled vocabularies are available at http://opendata.dsi.uminho.pt/terms/energy/.

4. Closing remarks and future work

The design of application profiles often involves several kinds of trade-offs. In many cases the choice is between levels of interoperability, and it is not uncommon to consider losing some semantic rigor in favour of the use of a property more used by the metadata community, potentially gaining interoperability. Or rather: to dispense with some interoperability in favour of semantic rigor. The option is not always easy and is frequently subject to controversy. In some cases we just have to give away something. The issue is: what can we afford to give away?

The SusCity project is an MIT Portugal project within the scope of smart cities. One of its tasks has to do with opening data generated in its scope, in particular data about energy consumption, energy performance of buildings and urban mobility. For this data to be opened as LOD and to maximise its interoperability, it is necessary to create MAPs. This article reported the process and the results of the development of a MAP for energy performance certificates of buildings that are issued by ADENE, a co-funder and partner of the project. The development of this MAP implied the development of a few more metadata artefacts: a metadata schema for the declaration of four new properties and a class, and four controlled vocabularies, three of which serve as constraints to the range of some properties. The fourth controlled vocabulary, on

measures to improve the energy performance of buildings, was encoded in SKOS on the request of ADENE.

These developments have forced some design decisions that are relevant to the metadata community. We chose to have only one class (the `EnergyPerformanceCertificate`), which is the domain of all properties of the MAP. The use of one class results from a balance between power of description and simplicity: in order to facilitate the use and reuse of data, we have given up some power of description in exchange for more simplicity. Another decision we made has to do with rigour in the choice of properties. We consider imperative to make all efforts to respect the constraints defined in the schemas of the properties and this is what we did. However, there were cases where we had to leave something behind, as it happened with the compatibility of ranges. In some situations, our option was to dispense compatibility of ranges to be able to cope with the other criteria and use just one property per attribute. Another design option relates to the selection of properties when more that one respected our needs. In the face of two or more properties, which one should we choose? Our choice fell on the following priority criteria: 1) property in a schema corresponding to some standard or some recommendation of some relevant entity in the field of application (energy performance) or in LOD; 2) property widely used by the metadata community; or, finally, 3) property defined in a formal schema and used by at least two datasets. We only created new properties when none of these options were possible.

The creation of the metadata schema and controlled vocabularies involved the use of Wikidata, so all new terms (RDFS classes and properties and SKOS concepts) are related to Wikidata terms through specific OWL properties. The use of Wikidata to "declare" the terms was another design decision that we find relevant for evolution/scrutinity, availability and interoperability purposes. The RDFS and SKOS files also hold the vocabularies structure, leaving all created Wikidata terms free of any constraints. This facilitates the reuse of terms across different schemas while safeguarding minimal semantics.

Besides being available as RDFS and SKOS files, the vocabularies will also be stored in a triplestore together with other SusCity RDF data on energy performance and mobility. Future work includes the finalization of this process with the inclusion of the metadata schema in LOV and the design of new application profiles regarding energy consumption and mobility.

Acknowledgements

We would like to thank the following MSc students who participated in the prior study referred in the Introduction of this article: Diana Postolaki, Henrique Santos Guimarães and Ivo Filipe Marques Menezes.

We would like to thank Joana Fernandes (ADENE) for her interest, for the fruitful discussions we had, and for making available the human-readable version of the controlled vocabulary for measures to improve the energy performance of buildings.

Lastly, we would like to thank the reviewers of this article who, with their interest and commitment, contributed to the clarification of several aspects of the article and improvement of others.

This work has been supported by FCT – Fundação para a Ciência e Tecnologia in the scope of the project FCT/MITP-TB/CS/0026/2013.

Namespaces

dct - http://purl.org/dc/terms/

dbpedia-owl - http://dbpedia.org/ontology/

ec - http://opendata.dsi.uminho.pt/terms/energy/EPCterms/

osadm - http://data.ordnancesurvey.co.uk/ontology/admingeo/

owl - http://www.w3.org/2002/07/owl#

rdf - http://www.w3.org/1999/02/22-rdf-syntax-ns#

rdfs - http://www.w3.org/2000/01/rdf-schema#

skos - http://www.w3.org/2004/02/skos/core#

vcard - http://www.w3.org/2006/vcard/

xsd: http://www.w3.org/2001/XMLSchema#

References

ADENE. (2017). Quem somos | ADENE - Agência para a Energia. Retrieved 14 June 2017, from http://www.adene.pt/quem-somos

Aelenei, L., Ferreira, A., Monteiro, C. S., Gomes, R., Gonçalves, H., Camelo, S., & Silva, C. (2016). Smart City: A Systematic Approach towards a Sustainable Urban Transformation. *Proceedings of the 4th International Conference on Solar Heating and Cooling for Buildings and Industry (SHC 2015)*, *91*(Supplement C), 970–979. https://doi.org/10.1016/j.egypro.2016.06.264

Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart Cities in Europe. *Journal of Urban Technology*, *18*(2), 65–82. https://doi.org/10.1080/10630732.2011.601117

Costa, C., & Santos, M. Y. (2015). Improving cities sustainability through the use of data mining in a context of big city data. In *The 2015 International Conference of Data Mining and Knowledge Engineering* (Vol. 1, pp. 320–325). IAENG.

Coyle, K., & Baker, T. (2009, May 18). Guidelines for Dublin Core Application Profiles. Retrieved 28 June 2010, from http://dublincore.org/documents/2009/05/18/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 K. Eckert & M. Foulonneau (Eds.) (pp. 90–103). Lisbon, Portugal: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81

European Parliament and the Council of the European Union. Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings, Official Journal L 001 , 04/01/2003 P. 0065 - 0071 § (2003).

MIT Portugal. (2017). Retrieved 14 June 2017, from https://www.mitportugal.org/

Odyssee-Mure. (2012). *Energy Efficiency Trends in Buildings in the EU Lessons from the ODYSSEE MURE project*. Retrieved from https://www.nahb.org/ /media/Sites/NAHB/LMA/FileUploads/35601-1-REPORT_local_20150318115955.ashx?la=en

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings*, *40*(3), 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Sozer, H. (2010). Improving energy efficiency through the design of the building envelope. *Building and Environment*, *45*(12), 2581–2593. https://doi.org/10.1016/j.buildenv.2010.05.004

SUSCITY. (2016). SUSCITY – An MIT Portugal project. Retrieved 14 June 2017, from http://groups.ist.utl.pt/suscity-project/home/

Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2017). Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web*, *8*(3), 437–452. https://doi.org/10.3233/SW-160213

Expanding the Institutional Repository Mission: Innovating with Linked Data for NASA Digital Curation

Presentation

Adrienne Hieb
NASA Goddard Library
Cadence Group, Inc.
adrienne.m.hieb@nasa.gov

Matthew Pearson
NASA Goddard Library
ZAI, Inc.
matthew.m.pearson@nasa.gov

Mitchell Shelton
NASA Goddard Library
ZAI, Inc.
mitchell.shelton@nasa.gov

Abstract

The NASA Goddard Space Flight Center Institutional Repository (GSFCIR) manages, preserves, tracks, and provides access to the Center's digital collections and research output. As GSFCIR moves to an entirely RDF-based platform, the Goddard Library is taking this opportunity to leverage linked data's capabilities to enhance digital curation efforts, particularly in the area of adding value to digital collections.

Objects in GSFCIR's existing collections have little inter-relation through back-end metadata or front-end interfaces. As representatives of the research and knowledge output of Goddard, these collections and digital objects do have a common thread among them: NASA missions. Current repository cross-collection searching allows for discovery of some of these connections; however, it is often frustrated by variant names and it does not support a variety of common search behaviors.

Historically, NASA mission information has not been maintained in any single, accessible authority. To both achieve its goal of creating better connections in GSFCIR and to provide a valuable resource to present and future NASA communities, the Goddard Library is producing a linked data thesaurus of NASA mission names, including equivalence, hierarchical, and associative relationships.

This presentation will focus on how the Library established the need for a NASA-focused linked data missions thesaurus, the careful process of domain analysis and vocabulary development, and its role in aiding future digital curation efforts as GSFCIR grows with new collections.

Towards a BIBFRAME Implementation: The bibliotek-o Framework

Jason Kovari
Cornell University, USA
jak473@cornell.edu

Steven Folsom
Cornell University, USA
sf433@cornell.edu

Rebecca Younes
Cornell University, USA
rebecca.younes@cornell.edu

Abstract

bibliotek-o is a framework for modeling bibliographic metadata as linked data, consisting of the BIBFRAME ontology at its core. This paper presents the background and motivation behind the bibliotek-o framework, including an overview of the model, ontology principles and best practices guiding its development, a description of aligned tooling under development, and a report on the project's status and outputs. A small sample of discrete ontology design patterns in which bibliotek-o deviates from BIBFRAME is provided to demonstrate motivations and modeling principles. Our goal is to illustrate the strengths of BIBFRAME, while suggesting areas where BIBFRAME should evolve to a more streamlined and expressive model, such as in the treatment of Activities and Content/Carrier/Media Types. We aim to encourage feedback and community engagement in ongoing development of the framework outlined in this paper.

Keywords: bibliotek-o; BIBFRAME; bibliographic metadata; data modeling; linked data; ontology development

1. Introduction

bibliotek-o is a framework for modeling bibliographic metadata as linked data based on the BIBFRAME ontology (http://id.loc.gov/ontologies/bibframe.rdf), consisting of the BIBFRAME ontology at its core; the bibliotek-o ontology, which both extends and provides alternative models to BIBFRAME; defined fragments of external ontologies, both within and outside the bibliographic domain; and an application profile specifying the recommended implementation of these ontologies. Our aim is to illustrate the strengths of BIBFRAME, while suggesting areas where BIBFRAME should evolve to both simplify the model and be more expressive.

A joint effort of the Andrew W. Mellon Foundation funded Linked Data for Libraries Labs (LD4L Labs: http://ld4l.org/ld4l-labs/) and Linked Data for Production (LD4P: http://ld4p.org) projects, this work represents significant effort by a large group of colleagues from Columbia, Cornell, Harvard, Princeton and Stanford Universities as well as the Library of Congress (hereafter "LD4 Ontology Group"). LD4L Labs and LD4P are complementary efforts in support of tool development, RDF metadata production, community engagement and ontology development in the library realm.

The broader library community should determine implementation and evolution of BIBFRAME through experimentation and analysis, facilitated by transparent processes; bibliotek-o represents the LD4L Labs and LD4P approach to this analysis. bibliotek-o is not intended to replace or compete with BIBFRAME. Instead, it seeks to expand and provide proofs-of-concept for alternative modeling patterns to BIBFRAME in select modeling areas where the LD4 Ontology Group recommends more expressive and/or streamlined models without losing semantics.

Note that both BIBFRAME and bibliotek-o represent "core" bibliographic descriptive practice. Within the LD4P project, community-based domain ontology extensions are being developed concurrently with bibliotek-o development; these efforts at times inform and are informed by bibliotek-o, but represent a separate development stream.

This paper presents the background and motivation behind the bibliotek-o framework recommendation, the ontology principles and best practices guiding its development, an overview of the model, a description of tooling under development in support of analysis and testing of the model, and a report on the project's status and outputs. To demonstrate motivations and modeling principles, a small sample of discrete ontology design patterns in which bibliotek-o deviates from BIBFRAME is provided.

2. Background and Motivation

Principles behind bibliotek-o were introduced to the DCMI community during the 2016 conference in Copenhagen, DK. Focusing on modeling principles and select extension efforts, the presentation by Folsom and Kovari (2016) provided a very high-level overview of ontology efforts within the space of LD4L Labs and LD4P. Since then, efforts related to bibliographic data modeling have progressed substantially into a framework including multiple ontologies alongside an in-development application profile.

Development on bibliotek-o began with an assessment of BIBFRAME 2.0, focusing on the question of alignment with recommendations written by Rob Sanderson (2015) in review of BIBFRAME 1.0. Concurrent to Sanderson's development of the 2015 report, the Linked Data for Libraries (2014-2016) Ontology Group deemed BIBFRAME 1.0 insufficient for the description of library resources, see: https://ld4l.org/ld4l-2014/overview; as a workaround, the group developed a temporary ontology, modeled and implemented solely for achieving the goals of the LD4L 2014-2016 project and illustrating the concrete implementation of its recommendations. With the start of LD4L Labs and LD4P in 2016, the newly formed LD4 Ontology Group wished to deprecate the LD4L 2014-2016 ontology, hoping to implement BIBFRAME 2.0 wholesale under the assumption of full alignment with Sanderson (2015) and other recommendations. This was not the case.

During the alignment analysis, the LD4 Ontology Group noted significant improvements over BIBFRAME 1.0; however, there remained areas of "core" description not provisioned in BIBFRAME 2.0, as well as modeling decisions made by the BIBFRAME architects with which the LD4 Ontology Group disagreed. Thus began the development of the bibliotek-o framework. At the framework's core is BIBFRAME; bibliotek-o builds upon BIBFRAME and cannot be implemented without select BIBFRAME patterns, alongside a number of other external ontology fragments ("target" ontologies).

The goal of the LD4 Ontology Group is to provide an ontology as the basis for an RDF cataloging tool and a MARC-to-RDF converter, and to use the resulting RDF instance data for analysis and for testing hypotheses. We hope that bibliotek-o will provide for a richer model to represent bibliographic data than using BIBFRAME alone; however, the purpose of this development is to provide the ability to analyze both BIBFRAME and the bibliotek-o framework to determine whether either provides models that adequately balance cataloging use cases and users' discovery needs. The metrics for evaluation of either framework remain undeveloped.

The LD4 Ontology Group wishes to engage the community in development of bibliotek-o as a method of analyzing BIBFRAME. To do so, we demonstrate alternative ontology design patterns that we believe more closely align with linked data principles; further, we believe that these patterns yield more accurate and queryable models than the corresponding BIBFRAME patterns. Through this development, we aim to foster a dialogue with the community pertaining to alternative models for consideration as BIBFRAME evolves in future versions. If the community decides bibliotek-o modeling more successfully addresses cataloging use cases and user querying expectations, we hope that bibliotek-o's extensions and alternative design patterns will converge with BIBFRAME in future releases. Although Library of Congress is represented on the LD4 Ontology Group and was deeply involved in discussions around bibliotek-o development, there is no defined plan regarding convergence of BIBFRAME and bibliotek-o.

The LD4 Ontology Group does not anticipate supporting bibliotek-o in perpetuity; instead, our aim is to encourage community discussion around bibliographic modeling questions and deprecate bibliotek-o as the community decides upon modeling patterns. During this period of transition, LD4P partners intend to create linked data for use by the community using the bibliotek-o model as well as the BIBFRAME model. This will assure the community that our data model is safe to implement and our instance data is safe to consume even in the near term if so desired.

It is important to note that not all LD4P members are using the bibliotek-o framework. Library of Congress and Stanford University are building their LD4P projects from BIBFRAME without the extension and alternative patterns provisioned as part of bibliotek-o. This further affords assessment of the two models and resultant data; in addition, this divergence encourages future investigation of the effects of multiple "core" ontologies and frameworks in both cooperative cataloging and discovery environments, an issue that will absolutely arise in a shift to linked data production and consumption. As a collaboration, LD4L/LD4P expect to share data with each other regardless of the models we use. This underscores the need to consider strategies to account for the fact that there are multiple existing core ontologies already in use in the wider library community (e.g., BIBFRAME, RDA Elements, Schema.org, CIDOC CRM).

3. Development Process

The bibliotek-o ontology underwent a formal development process in April-December 2016. Self-selected representatives from each institution, including Library of Congress, formed an ontology development group to bring proposals back to the full ontology group.

The development subgroup began with a detailed alignment of BIBFRAME to the LD4L 2014-2016 ontology; this revealed areas where BIBFRAME 2.0 had not implemented those recommendations, as well as areas that neither ontology modeled adequately. After prioritizing these areas (time and resource constraints required that many would have to be left for future work), the group engaged in frequent and regular analysis and discussions that resulted in a series of recommendations documenting new requests to Library of Congress for BIBFRAME modifications, including motivations, design pattern diagrams, and complete enumeration of affected terms.

Throughout this process, our default stance was to ask Library of Congress for changes in BIBFRAME before implementing in the bibliotek-o ontology. These requests included deprecating BIBFRAME classes and predicates and adopting those from more established ontologies, addition of semantics via OWL axioms, new modeling patterns that we deemed more expressive, and miscellaneous small changes and corrections. BIBFRAME evolved in some substantial ways in response to these requests: a few properties were changed from datatype to object properties (e.g. bf:projection, bf:contentAccessibility); new classes were added to correspond with (as domains or ranges of) new object properties; inverse properties were declared for most BIBFRAME object properties; properties were defined as symmetric where appropriate; domains and ranges of some properties were removed to broaden their applicability; and the BIBFRAME agent classes were declared subclasses of the corresponding FOAF properties. There were also a number of smaller changes made in response to the LD4 Ontology Group recommendations.

Where Library of Congress chose not to follow LD4 Ontology Group recommendations, the group made decisions about whether to defer to the existing BIBFRAME implementation or to proceed with an implementation in its own namespace. We took the former course for design patterns that we deemed not sufficiently important to warrant deviation from BIBFRAME, and the latter for those we considered of crucial importance; these resulted in the bibliotek-o ontology.

4. Design Principles

bibliotek-o differs from BIBFRAME in both modeling principles and modeling patterns that derive from those principles. The principles guiding the development of bibliotek-o include:

- Reuse and align with existing external vocabularies to promote data exchange and interoperability.
- Conversely, define terms broadly enough for reuse by external data sources.
- Use OWL axioms and RDF constructs such as domain and range in moderation to provide expressivity without overly constraining the ontology and the data it can model.
- Prefer object properties, structured data, and controlled vocabularies over unstructured literals.
- Prefer the simplest, most streamlined model capable of faithfully representing the data; adopt a single method of expressing a relationship or attribute in order to minimize query paths.
- Prefer atomic over composite data representation, e.g. bf:MovingImage and bf:Cartography instead of a class ex:CartographicMovingImage, which is an RDA content types.

Differences in modeling patterns are described in the following sections.

5. Overview of the bibliotek-o Framework

Figure 1 provides a high-level visualization of the bibliotek-o framework.

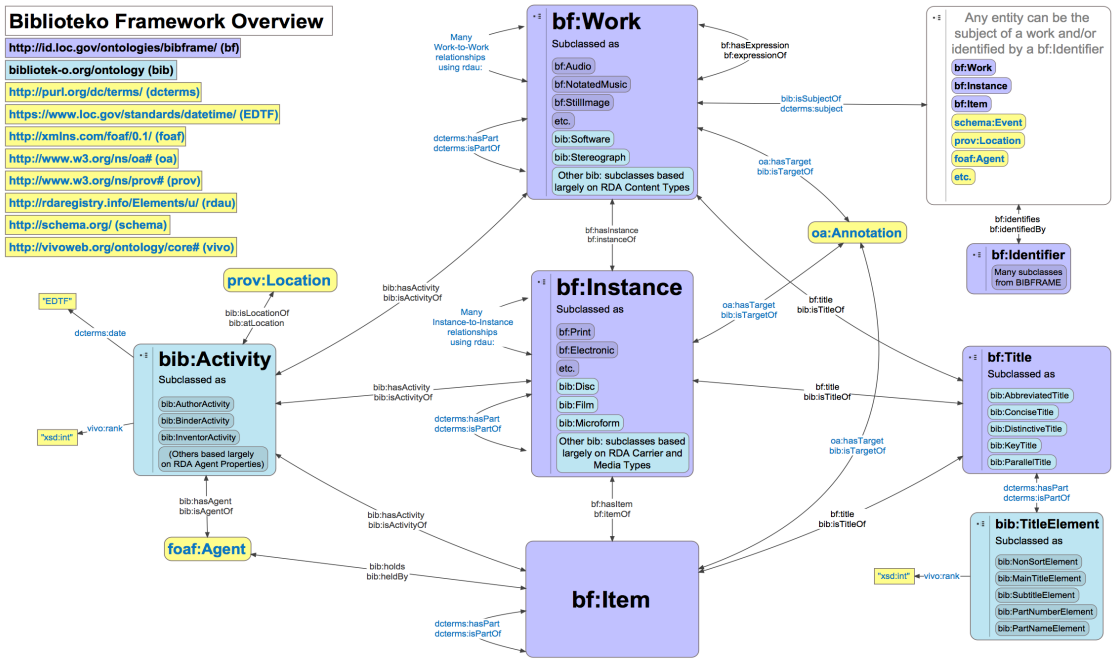


FIG. 1. bibliotek-o Framework Overview

As stated earlier, BIBFRAME is the base ontology within the bibliotek-o framework; the bibliotek-o ontology and other target ontology fragments provide additional semantics for core bibliographic descriptive practice. While these ontologies expand well beyond what could be considered a "core description" or "core record", "core" in this context means provisioning for general cataloging descriptive practice. Neither BIBFRAME nor bibliotek-o are intended to sufficiently provision for specialized cataloging practice. For instance, neither ontology would

enable a rare materials cataloger to sufficiently describe resources without either substantial loss of granularity or omission of crucial data points.

Again, one foundational principle of bibliotek-o is the reuse of existing classes and properties where semantics align and there has been substantial community development and adoption; this is outlined in the Reuse and Alignment Principle by LD4L Labs / LD4P Ontology Group (2016). For instance, dcterms:subject has deep usage in the library domain; we thus chose to reuse the dcterms predicate rather than implementing bf:subject. BIBFRAME does not reuse existing ontology fragments or patterns, instead opting to mint new terms and in a limited set of cases assert subclass relationships to external classes. The reasons for this are beyond the scope of this paper; see Library of Congress (2017).

Other high-level areas of divergence between BIBFRAME and bibliotek-o are described in design pattern documents housed in the bibliotek-o wiki (https://wiki.duraspace.org/x/H5TBB); these patterns do not represent a full discussion of the points of deviation, which is currently in development and will be available at the same location once complete. Examples of these divergent patterns are provided in the following section.

Because bibliotek-o extends beyond BIBFRAME and in many cases adopts richer and more expressive models, transformation from bibliotek-o to BIBFRAME is lossy. That said, part of this effort is to determine whether more granular modeling with precise semantics has a positive effect on data querying and thus user discovery (though the current LD4L Labs and LD4P projects do not include a customized discovery environment). As data in both BIBFRAME and bibliotek-o is made available and consumed by both ourselves and others, we can evaluate whether bibliotek-o's streamlined and more resource-centric model meets our needs.

6. Legacy Data and Object vs. Datatype Properties

A significant challenge for an ontology representing bibliographic metadata is to bridge the competing demands to both migrate the existing highly detailed and nuanced data, and prepare for a future of original cataloging in RDF that captures data in meaningful and useful ways with a real-world orientation.

While BIBFRAME is often oriented towards preserving existing MARC data in its current format as string literals, the LD4 Ontology Group has emphasized modeling for the future without loss of legacy data. bibliotek-o thus replaces many BIBFRAME datatype properties with object properties, using the former only for data which is truly unstructured by nature, such as bf:responsibilityStatement and bf:provisionActivityStatement (though this goal has not been fully realized in the current version of bibliotek-o due to time constraints). Where tools do not currently exist to meaningfully parse and structure the legacy data, the bibliotek-o framework recommends using object properties and creating resources to hold this data in a generic datatype property such as rdf:value, and defines a custom datatype to flag such data for future processing. This approach prevents distorting the model in order to accommodate unstructured data, while nevertheless preserving that data for future integration into the model.

7. Modeling Patterns

The sections below demonstrate select design patterns where bibliotek-o offers alternative modeling from BIBFRAME and one pattern where we demonstrate an extension pattern for BIBFRAME. The LD4 Ontology Group believes that these patterns provision for better descriptive practice. These patterns, and others, are documented in significantly more detail in the bibliotek-o wiki.

7.1 Activities

BIBFRAME makes a distinction between Provision activities (Distribution, Manufacture, Publication and Production) and a Contribution activity, the former applying to Instances and the latter to Works. While the LD4 Ontology Group understood the reasoning behind this distinction, we anticipated that the profiling of these classes would be closely aligned and that the sharp division between Instance- and Work-related activities might not be fully sustainable. We believe that the bifurcation is unnecessary and results in overly complex query paths, which should be tested once datasets are available.

As a result, we defined a general Activity pattern that provisions for explicit roles through subclassing of the bib:Activity class which links Agents to Works, Instances and Items, eliminating the distinction between provisions and contributions. This pattern also allows potential extension to Activity relationships with other types of resources, such as events. This basic design pattern is adopted in other ontologies, such as the Schema.org Action class (http://schema.org/Action) and the CIDOC CRM Activity class (http://www.cidoc-crm.org/Entity/e7-activity/version-6.2). While there was an attempt to reuse related terms from these other models, the LD4 Ontology Group agreed to mint new bibliotek-o terms until we had more experience with the pattern's semantics and our related use cases. Alignment with related external models was identified as future work. The diagram below provides a visualization of the bibliotek-o model:

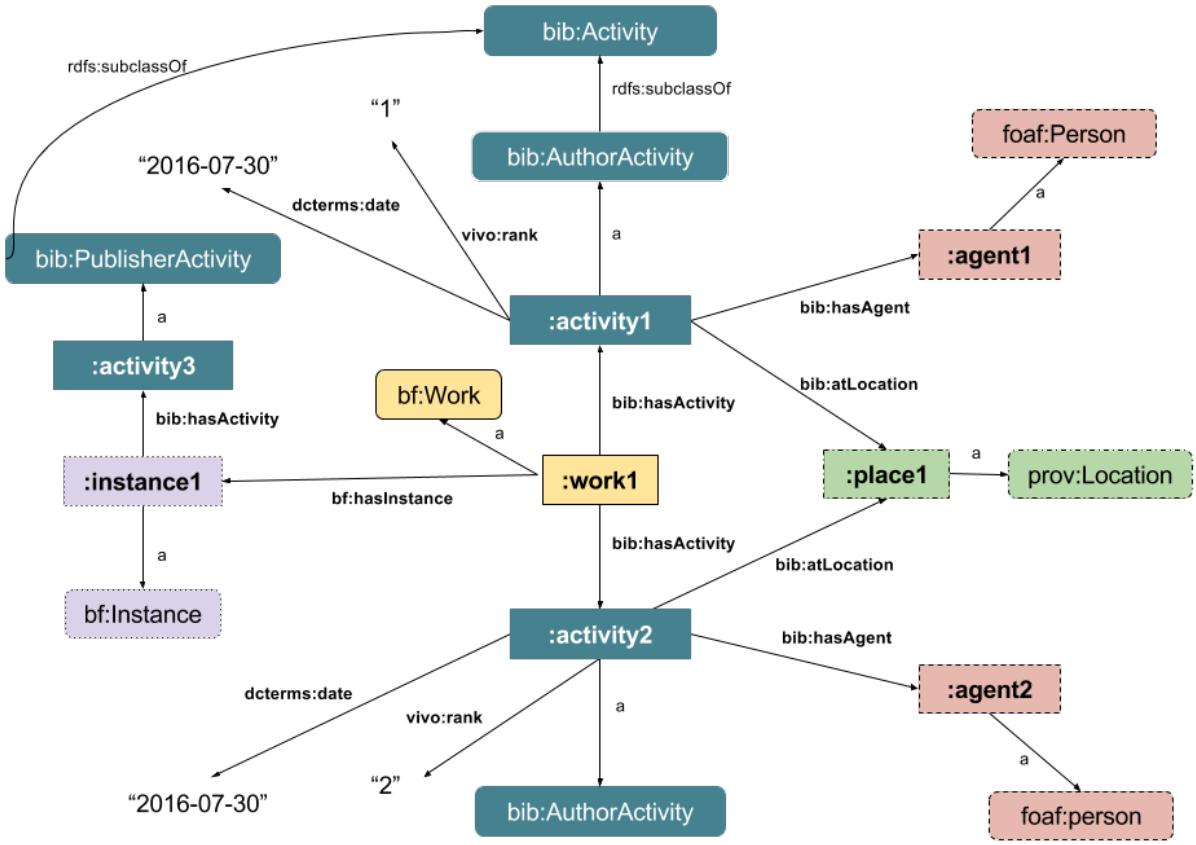


FIG. 2. bibliotek-o Activity Model

7.2 Content/carrier/media

BIBFRAME asserts content types, carrier types and media types by establishing bf:content/bf:Content, bf:carrier/bf:Carrier, bf:media/bf:Media patterns. This modeling creates two potential means for stating the same thing: through subclassing bf:Work, bf:Instance and bf:Item, or through the property/class pattern referenced above. The LD4 Ontology Group believes that this should be avoided, as it demands that consumers of this data perform more complex queries to identify all things of a certain type (see the principles enumerated above); further, it diverges from the standard linked data practice of using rdf:type to declare that a resource is a particular kind of thing.

Rather than using the BIBFRAME multi-path pattern, bibliotek-o models content types, carrier types and media types as rdf:type assertions directly on the resource. After some testing through the creation and use of data according to defined subclasses, we may find the need to reconsider the class hierarchies and related definitions.

Library cataloging practice has a long history of capturing content type, carrier type and media type. Capturing types corresponding to content and carrier directly on Works, Instances and Items through the use of rdf:type can still be interpreted as an RDA implementation pattern because it captures content/carrier/media information about library resources; thus, this pattern does follow the RDA content standard. Note, as shown in the example below, entities can have multiple type assertions.

The following diagram visualizes the bibliotek-o design pattern:

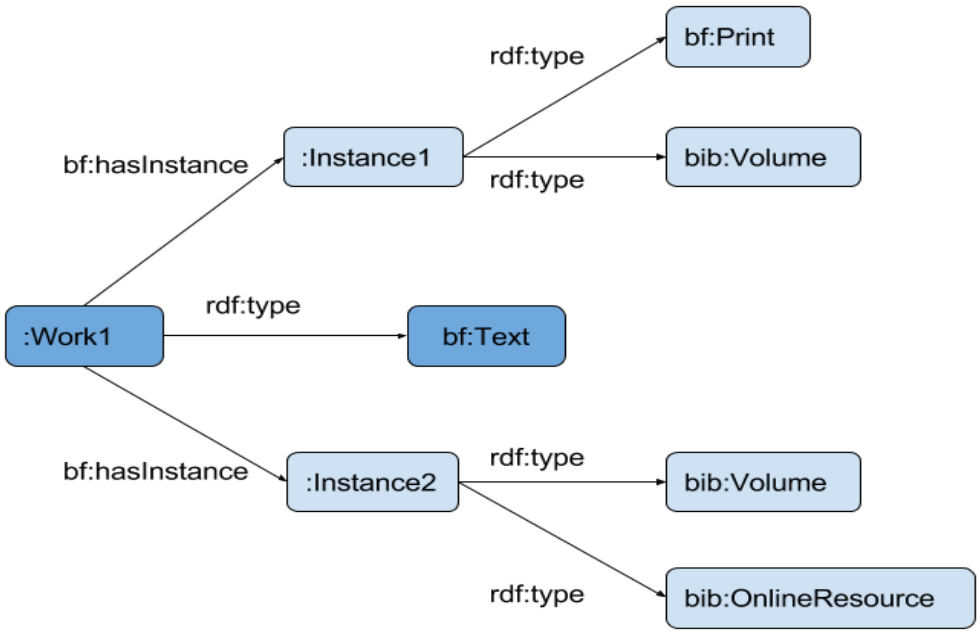


FIG. 3. bibliotek-o Content/Carrier/Media Model

8. Status and Implementation

As of the final submission of this paper (September 2017), the following technical outputs have been achieved:

- Publication of version 1.0 of the bibliotek-o ontology:
  - OWL file: http://bibliotek-o.org/ontology.owl
  - Human-readable documentation: http://bibliotek-o.org/ontology.html
  - GitHub repository: https://github.com/ld4l-labs/bibliotek-o/tree/v1.0.1

- Specification of fragments from other ontologies, including the core BIBFRAME ontology, that are part of the bibliotek-o framework
- Recommendation documents for bibliotek-o principles and design patterns, published: https://wiki.duraspace.org/x/H5TBB

The following technical outputs are in progress, and are intended to facilitate implementation of the bibliotek-o framework in tooling (see section on Tooling):

- An application profile specifying expected use of classes and properties, to ensure consistent application of terms across data sets
- A mapping from MARC to bibliotek-o
- A listing of all BIBFRAME classes and properties not used in bibliotek-o

9. Tooling

Based on the bibliotek-o framework, tools for both original RDF metadata creation and conversion of existing metadata are currently under construction to facilitate further analysis of and experimentation with the framework.

In a joint effort between LD4L Labs and LD4P to begin creating RDF natively within the bibliotek-o framework, we have begun the development of VitroLib (https://github.com/ld4l-labs/vitrolib), an ontology-based cataloging editor that enables manual cataloging and is built on Vitro (https://github.com/vivo-project/Vitro), a generalized semantic web ontology and instance editor with customizable browsing. As part of this work, usability testing is being performed with catalogers to better understand how VitroLib can be customized to conform to cataloger needs and expectations. Considerable work is being devoted to ensuring that VitroLib cataloging forms not only allow for lookups of locally created data, but also provision for linking to external data in much the way that we reuse bibliographic records and authority data in current cataloging workflows. As ontology extensions become available from the LD4P domain projects, they will be incorporated into VitroLib implementations to achieve richer descriptions of particular domains. Additionally, we will begin testing customization of CEDAR (https://metadatacenter.org/tools-training/cedar-metadata-tools) for bibliotek-o during Fall 2017, customization will be based on the same in-development application profile as VitroLib.

The converter is an open source, flexible, community-extensible tool for the conversion of conventionally-formatted bibliographic metadata to RDF and linked data (https://github.com/ld4l-labs/bib2lod). In its primary implementation, it will convert core bibliographic metadata in MARC to RDF expressed in the bibliotek-o framework, based on mappings under development by the LD4 Ontology Group. It is architected for extensibility, allowing implementations that accommodate both other metadata input formats (e.g. FGDC (http://www.fgdc.gov/metadata/fgdc-std-001-1998.dtd) and CSV) and output formats that extend the bibliotek-o framework, such as the LD4P domain ontologies; in fact, an extension converting FGDC data to bibliotek-o plus the cartographic extension has already been developed. The converter is capable of converting catalog metadata at scale, allowing analysis and evaluation of large amounts of linked data built on the bibliotek-o framework.

10. Community Engagement

The LD4 Ontology Group encourages feedback on any fragments and modeling patterns represented in bibliotek-o. Documented in GitHub (https://github.com/ld4l-labs/bibliotek-o), we encourage the submission of GitHub issues to raise concerns about particular classes, properties and patterns; and pull requests to suggest actual changes to the ontology file and related documentation.

While in-depth ontology development for bibliotek-o concluded December 2016, development of extensions and application profiles is intended to continue, based on both LD4P domain extension work and community input. The bibliotek-o versioning protocols and change and release management process have been published in the GitHub repository (https://github.com/ld4l-labs/bibliotek-o/); these processes will ensure transparency of ontology development moving forward.

11. Select Outstanding Issues and Future Work

While we cannot address all outstanding questions adequately in this paper, there remain a number of issues in the bibliotek-o framework as well as in BIBFRAME that the LD4 Ontology Group would like to begin addressing with the community. These include, though are not limited to:

- Should BIBFRAME, bibliotek-o or other ontologies be made more modular? Is there a benefit to the broader (library and non-library) community if definable ontology patterns are managed and hosted in namespaces distinct from the core ontologies? For example, some of the LD4P domain extension ontology groups have surfaced common needs for provenance modeling; rather than including a provenance model in any one extension, should it be included in a BIBFRAME or bibliotek-o core, or hosted in a namespace of its own so that it can be implemented independently of a core bibliographic ontology?
- The LD4 Ontology Group identified a number of BIBFRAME models (e.g., Awards, Administrative Metadata, Degrees, Form/Genre, Serials and Multi-parts, etc.) which require further work before submitting recommendations to BIBFRAME or implementing in bibliotek-o. We do not regard the data model as complete without the inclusion of these models.
- In addition to alternative modeling patterns, the LD4 Ontology Group has considered areas of data representation that BIBFRAME, as well as traditional data formats and schema, have not addressed. One area of extensive exploration was Attribution: the complex web of relationships (over 20 uses cases were identified) that may occur between a bibliographic resource and attributed agents, such as: pseudonyms, ghostwriters, name changes, deliberate misattribution, etc.; note: relationships between identities and agents were not within the scope of this research. While existing metadata does not express such relationships, we view it as a fruitful area for future efforts to expand beyond existing descriptive domains.
- Methods and metrics for analysis and evaluation of the bibliotek-o framework based on the RDF generated during the project by original cataloging and bulk conversion, as well as comparison with BIBFRAME implementation data.

Acknowledgements

This work is supported by the Andrew W. Mellon Foundation. The authors would like to thank the Mellon Foundation as well as the numerous colleagues who contributed to this development from Columbia, Cornell, Harvard, Princeton and Stanford Universities as well as the Library of Congress.

References

Folsom, Steven, and Jason Kovari. (2016). Ontology Assessment and Extension: a case study from LD4L and BIBFRAME. Retrieved May 25, 2017, from http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/433/505.

LD4L Labs / LD4P Ontology Group. (2016). Reuse and Alignment Principle. Retrieved May 26, 2017, from https://github.com/LD4L Labs/bibliotek-o/blob/develop/doc/principles/bibliotek-o_principle_reuse_201612.md.

Library of Congress. (2017). BIBFRAME frequently asked questions. Retrieved May 26, 2017, from

https://www.loc.gov/bibframe/faqs/#q07.
Sanderson, Robert. (2015). Analysis of the BIBFRAME Ontology for Linked Data Best Practices. Retrieved May 25, 2017, from https://goo.gl/KRiuTt.

Sustainability and Preservation

Applying the Levels of Conceptual Interoperability Model to a Digital Library Ecosystem – A Case Study

Charlotte Kostelic
Library of Congress, USA
ckos@loc.gov

Abstract

This paper applies the Levels of Conceptual Interoperability Model to a case study of two cultural heritage institutions with disparate but related collections in an effort to define a maturity model for interoperability between presentations of digitized cultural heritage materials. The Levels of Conceptual Interoperability Model (LCIM) is a progressive model developed by Dr. Andreas Tolk within the field of Modeling and Simulation and systems engineering to be used in determining potential for interoperability between systems. This paper applies the LCIM through a descriptive model to a digital library ecosystem that includes digital collections, digital libraries, and meta-aggregators. This paper seeks to determine if this model is sufficient as a method of measuring the potential for interoperation between systems, metadata, and collections within a digital cultural heritage ecosystem. A maturity model for interoperability within a digital library ecosystem can aid metadata operations specialists in determining the potential for interoperability between systems and collections.

Keywords: Levels of Conceptual Interoperability Model; Linked Open Data; metadata interoperability; cultural heritage, maturity model

1. Introduction

The development of digital collections and digital libraries that has occurred over the last two decades has been characterized by rapid development of new technologies and standards. The adoption of new standards and implementation of new technologies can be limited by the human and monetary resources available within individual institutions. Providing access to digital surrogates and metadata for cultural heritage collections is inherently a project that seeks to transform access to collections. The examination of interoperability for digital collections is an exploration that seeks to enhance access to online collections by making connections between cultural heritage materials.

The case study for this paper explores how two institutions with metadata and digitization standards that have developed over time may be able to measure their potential for interoperability between their collections. The two institutions are the Library of Congress and Royal Collection Trust. These institutions have formed a partnership through which they will explore shared access to collections held at each institution related to Early American history and the Georgian period in the United Kingdom. In order to better serve their users – some of whom may wish to consult collections at both institutions – and to enrich public enjoyment of their collections, the Library of Congress and Royal Collection Trust have sought to explore the potential for interoperability between their collections.

In order to better understand interoperability between systems that provide access to digital collections, a clear definition of interoperability is needed. The Oxford English Dictionary defines interoperability as "the ability of two of more computer systems or pieces of software to exchange and subsequently make use of data," although this definition may not adequately express the complexity within the concept. Digital library ecosystems can include not only collections within an individual institution, but also related collections in other institutions or dispersed collections around the world. Online users of cultural heritage collections may view

materials around the world through individual institutions' websites or through aggregators. This complex ecosystem requires a more complex understanding of interoperability than the one provided by the Oxford English Dictionary.

An underlying idea of this paper is that cataloging and description in cultural heritage institutions are constantly evolving and improving with a goal of improving access. Within the Library of Congress and Royal Collection Trust this evolution has led to metadata being created, edited, enhanced, and amended as standards change and users request different methods of access. This evolution has resulted in rich metadata records in various formats, accessible in different ways, and presented in different locations. With these different records in many locations there is a need to determine how to employ data from all of them in order to enhance use of the data. Interoperability, access, and understanding are inherently linked in a digital library ecosystem because of need for composability of digital objects within an online presentation setting. Although the ability to compose digital objects online can allow for users to access collections without a level of interoperability there may be a low level of understanding of the collections. The Levels of Conceptual Interoperability Model can serve as a maturity model for interoperability within the digital cultural heritage sector that can help to measure access, interoperability, and understanding.

2. Methods of exploring interoperability

2.1 Traditional Methods of Developing Interoperability

In a digital cultural heritage ecosystem, interoperability can be explored through different approaches. Digital objects within this ecosystem can contain many different components including one or more metadata record and one or more digital surrogates for the object. In an online presentation of a digital collection both the digital surrogate and the metadata record for each digital object must conform to the expectations of the presentation system.

Rachel Heery (2004) focuses on interoperability through the lens of the Semantic Web and its potential for use in digital libraries. Her exploration of the take-up for new technologies in libraries highlights the difficulties that libraries face in deciding whether or not to implement new technologies. Heery highlights that the tradition of work with technology in libraries is characterized by collaboration, exchange, and consensus. Further, this tradition of collaboration and consensus can aid the library community when determining how to implement semantic web technologies. Heery outlines the difficulties associated with trying to implement semantic web technologies but asserts that this culture of collaboration and consensus can aid in implementation at a small scale through interworking within the library community. This means that implementation of semantic web technologies (an example of a way to level up in interoperability) could be done through community building, furthering the potential for interoperability.

Van de Sompel & Nelson (2015) present the opinions of two practitioners working for over fifteen years on efforts to improve interoperability within the realm of scholarly communication. While not specifically outlining a model for interoperability, Van de Sompel & Nelson (2015) highlight the distinctions between repository-centric and web-centric approaches to interoperability, noting the need to focus on web-centric approaches in order to meet user needs rather than repository-centric approaches that better serve machines. This emphasis on user-needs is significant as systems used to support scholarly communication – in the case of Van de Sompel & Nelson – or digital libraries – in the case of the Georgian Papers Programme – are ultimately only successful if they support research conducted by humans. Van de Sompel & Nelson focus on systems interaction – specifically REST/HATEOAS principles – in their definitions of interoperability rather than the metadata-specific approach to interoperability which I hope to define in this paper. Alipour-Hafezi et al. (2010) also highlight models of interoperability with an emphasis on how systems interact. Alipour-Hafezi note three models for interoperability in digital libraries: federated, harvesting, and gathering. Specific metadata schema and data formats that are

applicable to interoperability protocols are highlighted, but the article places less of an emphasis on the underlying data that is harvested by the protocols and more of an emphasis on how the systems interact in order to share and transfer data.

Metadata interoperability in a digital library environment has been explored in a tiered approach by Jian Qin and Marcia Lei Zeng (2016). Qin & Zeng (2016) organize these levels based on "the point at which interoperability efforts are initiated" meaning that a different level of interoperability may be possible depending on whether the choice to make metadata records interoperable was made before or after cataloging guidelines were created, repositories were chosen, or records were created. This model was first introduced by Marcia Lei Zeng in Chan & Zeng (2006) where three levels of interoperability – schema, record, and repository – are highlighted. Chan & Zeng note the definition of interoperability within digital libraries highlighted in Tennant (2001) which helpfully defines interoperability as the ability for users to use one search to recall objects from many databases without needing to search each collection individually. Tennant's (2001) basic definition is expanded by Chan & Zang through their three levels. These three levels are not mutually exclusive and each requires data manipulation at different times in order to develop metadata records that conform to the expectations of the system. At each of Qin and Zeng's three levels – schema, record, and repository – the methods of achieving interoperability are limited by the need to achieve conformity using legacy metadata. This approach is both limiting – because understanding the metadata is limited to the description that already exists rather than future cataloging projects – and practical – because these approaches allow librarians and information professionals to use their existing data rather than use resources to create new data.

Nilsson et al. (2009) also published a set of levels that mark interoperability, although in this case the levels specifically related to compatibility with Dublin Core metadata within a specific application or specification. This "ladder of interoperability" contains four levels at which an institution can measure their compatibility and the levels present simple questions that serve as tests for with which to measure compatibility. Nilsson (2010) further reviews approaches to metadata interoperability and presents harmonization as the method of achieving interoperability between distinct metadata specifications. The five key components of harmonization within Nilsson's (2010) thesis are: syntaxes for metadata exchange, semantics to interpret metadata correctly, abstract models for designing standards, vocabularies as carriers of meaning, and application profiles used to combine standards. Nilsson (2010) applies definitions of metadata and interoperability to define the concept of metadata interoperability as "the ability of two or more systems or components to exchange descriptive data about things, and to interpret the descriptive data that has been exchanged in a way that is consistent with the interpretation of the creator of the data." The inclusion of a consistent interpretation of the data is a significant part of this definition. Nilsson (2010) highlights a model defined in Haslhofer & Klas (2010) made up of four levels with which to define a metadata model. This four-level model allows for different levels of interoperability to be defined based on the case study.

The purpose of creating a maturity model for interoperability within a digital library environment or digital cultural heritage ecosystem is to highlight ways in which access to digital cultural heritage materials may be improved. Nilsson (2010) and Haslhofer & Klas (2010) and their abstract metadata models are more aspirational than Qin & Zang (2016) because of the focus on legacy data presented in Qin & Zang (2016). The models presented by Nilsson (2010) and Haslhofer & Klas (2010) as well as their definitions of syntax and semantics can be combined to present a more detailed maturity model. The highest level within these models may not be practical for some institutions but it is a goal to work toward in order to improve access to cultural heritage materials.

2.2 Levels of Conceptual Interoperability Model

Applying an existing model for interoperability to the digital cultural heritage ecosystem will allow for a structured and consistent approach to interoperability. The Levels of Conceptual

Interoperability Model (LCIM) was developed by Dr. Andreas Tolk within the Modeling and Simulation discipline of systems engineering in order to recommend the use of "rigorous engineering methods and principles and replace ad-hoc approaches" to the development of interoperability (Tolk et al. 2009). The framework, as described by Tolk et al. (2009), has both descriptive and prescriptive uses in systems engineering. This paper will give focus to the descriptive uses of this model and will recommend equivalent examples of each level that exist in current practices and recommendations for the future. Tolk et al. (2009) assert that the purpose of the LCIM descriptive model is to "depict or analyze the ability, properties, characteristics and the levels of conceptual interoperability of an existing system or system of systems…" Thus one can evaluate the interaction between two systems and inform users of current interoperability potential.

TABLE 1: Levels of Conceptual Interoperability Model – Descriptive Model (Tolk et al., 2009)

Level	Layer Name	Description of level
L6	Conceptual	Interoperating systems at this level are completely aware of each other's information, processes, contexts, and modeling assumptions.
L5	Dynamic	Interoperating systems are able to re-orient information production and consumption based on understood changes to meaning, due to changing contexts.
L4	Pragmatic	Interoperating systems will be aware of the context (system states and processes) and meaning of information being exchanged.
L3	Semantic	Interoperating systems are exchanging a set of terms that they can semantically parse.
L2	Syntactic	Have an agreed protocol to exchange the right forms of data in the right order, but the meaning of data elements is not established.
L1	Technical	Have technical connection(s) and can exchange data between systems
L0	No	NA

Tolk et al. (2009) associate the different levels of their model with different concepts that describe how systems interact. The lower levels are concerned with integrability – meaning the ability for systems' hardware and protocols to interact. This is less complex than the middle levels' concern with interoperability – meaning the ability for systems to exchange data. The highest levels are concerned with composability – meaning the ability for systems to consistently represent data in context. In some ways the LCIM is misnamed because, while interoperability is a goal, the overall goal is for composability. Interoperability is merely one aspect of composability.

The LCIM in its descriptive role serves as a maturity model for interoperability within modeling and simulation. This means that each subsequent level must fulfill the requirements of all levels preceding levels. Tolk et al. (2009) note that the purpose of this descriptive model is to "describe how existing systems are interoperating and what level of conceptual interoperability can be reached by user's specific approaches without prescription." A digital library-specific application of the LCIM will be outlined in more detail in Table 4. The outcome of descriptive role can be used to evaluate the interoperability of existing systems and inform the users of the current properties and capabilities of interoperability. The flexibility of the descriptive model also allows for it to be applied to fields outside of the modeling and simulation field that are also concerned with the ability for systems to interact, share data, and maintain context.

3. Case Study – Georgian Papers Programme

3.1 Georgian Papers Programme

Beginning in 2015, the Georgian Papers Programme is a collaborative project within the Royal Collection Trust that aims to transform scholarly access to and personal enjoyment of the papers

of the Hanoverian monarchs ("About Georgian Papers Online," 2017). Over the course of the five-year program the Georgian Papers will be cataloged and scanned by the Royal Archives and Royal Library – two divisions under the umbrella of the Royal Collection Trust – with regular batches of papers being made available online. This artificial collection of papers has been assembled through a series of accessions to the Royal Archives and includes the official and private papers of King George I, II, III, and IV as well as other members of the Royal Family and Royal Household of the United Kingdom from the 18th and early 19th century ("What's in the catalogue?," 2017). To the end of transforming access to collections, the Royal Collection Trust partnered with institutions in the United Kingdom and United States in order to sponsor academic research fellowships and technical exploration. The Library of Congress has partnered with the Georgian Papers Programme in order to add context to its own early American collections some of which, such as the George Washington Papers (1592-1943), were created or accumulated while America was still a colonial holding of the United Kingdom under King George III. The Library of Congress and Royal Collection Trust have jointly sponsored a National Digital Stewardship Resident who will explore the potential for interoperability between the Georgian Papers housed at the Royal Collection Trust and related early American manuscript, bibliographic, print, and map collections housed at the Library of Congress.

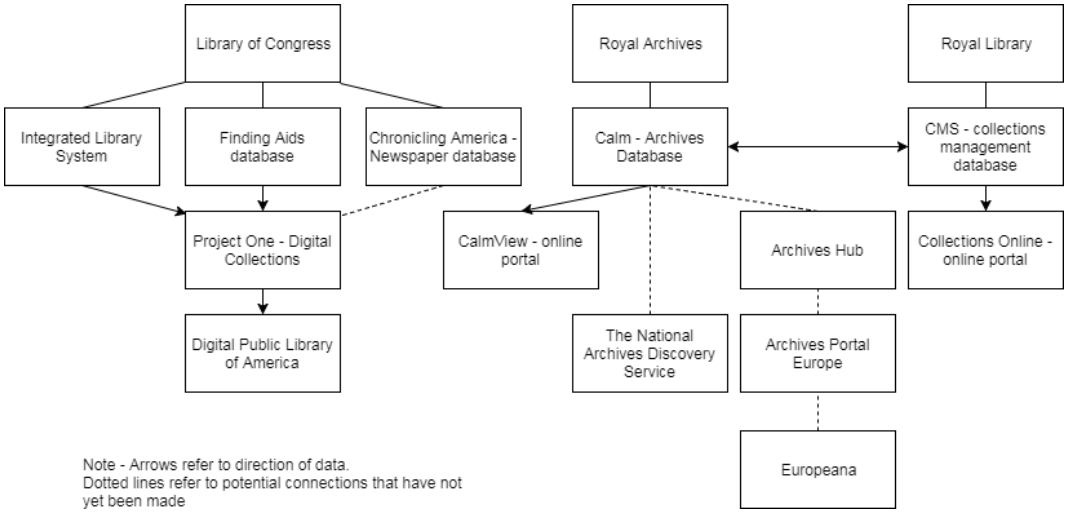
Both institutions have aspirations and plans to improve access to their collections online. Some collections are already accessible through online portals but further explorations could determine how these collections interact online with other collections and systems. While the impetus for the partnership between the two institutions was to make connections between collections related to the United Kingdom and its former colony, in order to make these connections there must be a way for users to examine materials across all collections. This exploration requires analyzing the individual institutions, their systems and metadata standards as well as the ways in which these systems and metadata standards may be interoperable.

While each of the partners involved in the project recognize the potential for improved access to their collections, there is not a universal definition of interoperability for its application to digital cultural heritage materials. Developing a definition of interoperability that will allow for the institutions to determine their potential will allow for partners to make it easier to determine what the goals are for the each of the partners individually as well as give them the potential to make decisions for collaborative projects related to interoperability of their collections. Each of the institutions is limited in their potential for interoperability based on practical limitations – such as levels of funding or staffing for further metadata enrichment projects – but the purpose of this exploration is to develop documentation and to describe the potential for interoperability rather than to try make recommendations for interoperable access for all of the collections.

3.2 Digital ecosystem for case study

The digital library ecosystem that is represented in this case study includes digital library systems, databases, and meta-aggregators. Figure 2 represents the digital ecosystem for the Georgian Papers Programme and the Library of Congress as it is explored through this interoperability project. Platforms that are not yet connected but may be in the future are noted with dotted lines. In particular, the Royal Archives has not yet established partnerships with Archives Hub or other meta-aggregators, but may establish these partnerships in the future. Data transfer is noted with arrows.

FIG. 2. Digital Ecosystem for Georgian Papers Programme



3.3 Current state of collections

In order to determine the potential of future interoperability between the collections, the current state of the collections, their metadata, and other components of their digital objects must be defined. Although there is an emphasis on interoperability between descriptive metadata for the collections, this analysis is concerned more broadly with interoperability between the collections and their components as well as the potential for interoperability between systems that deliver access to digital collections. Table 3 outlines the descriptive standards, syntaxes, and schemas used by the institutions to catalog their collections.

Table 3: Collection metadata schemas and syntaxes

Collection	Descriptive Standard	Schema	Syntax
George Washington Papers	Describing Archives: a Content Standard	EAD 2002 MARC21 MODS METS Project One Element Set	XML XML XML XML JSON
Benjamin Franklin Papers	Describing Archives: a Content Standard	EAD 2002 MARC21 MODS METS Project One Element Set	XML XML XML XML JSON
British Cartoon Prints	AACR2	MARC21 Dublin Core MODS Project One Element Set	XML XML XML JSON
American Revolutionary War Era Maps	AACR2	MARC21 Dublin Core MODS Project One Element Set	XML XML XML JSON
Early American Newspapers	AACR2	MARC21	XML
Georgian Papers (Archives)	ISAD(G)	EAD 2002 Calm Element Set	XML XML
Georgian Papers (Library)	N/A	RCT Element Set	[expressed in CSV]

The manuscript collections within the Georgian Papers Programme at the Royal Collection Trust are described using cataloging guidelines developed in-house. These guidelines are documented and available internally. The guidelines are an interpretation of ISAD(G) and the resulting records can be expressed in EAD 2002. The Royal Library collections have been described using a schema developed for the Royal Collection Trust to be used within its bespoke database. This schema captures key data for the objects within the library collections including creators, date created, physical descriptions, and other data points. The data can be exported from the database to CSV and Word documents using report formats created by internal developers but the metadata does not comply with any international standard descriptive practices or any metadata schema.

A select number of collections held by the Library of Congress have been selected for inclusion in this analysis due to their thematic similarities to the Georgian Papers including the George Washington Papers, Benjamin Franklin Papers, British Cartoon Prints, an assembled collection of American Revolutionary War era maps, and select early American newspapers. These collections have been cataloged and will all be accessible online by the end of 2017 although most collections are currently accessible online.

The collections are described using cataloging guidelines consistent with practices for each discipline. The George Washington and Benjamin Franklin Papers are both described at the object level in finding aids that are compliant with Describing Archives: a Content Standard (DACS). Select portions of the George Washington Papers have full-text transcriptions. These finding aids can be expressed in EAD 2002 and the data has also been mapped to MODS, METS, and MARC which can all be expressed in XML. The British Cartoon Prints and American Revolutionary War Era maps collections are both described at the item level in MARC records. These records have also been mapped to Dublin Core and MODS and are expressed in XML. All of the metadata records for the manuscript, print, and map collections have also been transformed into a data model for the Library of Congress Digital Library platform Project One. This data set is expressed in JSON. The Early American newspaper collections have minimal metadata records for their online presentation – containing only titles, dates, and page numbers – but have associated MARC records that describe the newspaper as a whole. These newspapers have been transcribed using optical character recognition (OCR) and are full-text searchable.

3.3 Current sharing capabilities and functions

Both the Library of Congress and the Royal Collection Trust currently have multiple online portals through which users are able to access collections. These systems have differing levels of data export capabilities. These systems serve as access portals for the various collections and some can also export data to specific formats. There are additional formats available for export directly from databases rather than from the online portals.

The Library of Congress collections are published through a finding aid database, an integrated library system that supports MARC records, and a digital library system that publishes digital collections. The integrated library system is able to export records to MARC and the finding aids database is able to export records to EAD 2002, MODS, and METS. These databases do not currently have connections to other systems outside of the Library of Congress but the data within these databases are transformed into records within the digital library system, Project One.

The digital library system, Project One, provides access to all digital collections with collection metadata mapped from existing descriptive metadata. Project One can be accessed through an online interface or via API. The Library of Congress is also a Content Hub for the Digital Public Library of America (DPLA) with the Project One API serving as the point from which data is shared with the meta-aggregator. The Project One element set has been mapped to the DPLA Metadata Application Profile for collections that have been provided to DPLA. Library of Congress collections that have been made accessible in DPLA can also be accessed using the DPLA API.

As separate divisions within the Royal Collection Trust, the Royal Archives and Royal Library each have separate online access platforms where Georgian Papers collections have been made accessible. Royal Library collections, including the Georgian collection, can be made accessible through Collections Online, the Royal Collection Trust online collections portal. This platform provides access to collections related to all of the different divisions of the Royal Collection Trust. The Royal Archives collections are currently accessible through a public viewer – CalmView – that is published from their database. These public access portals do not allow users to export data to any standardized formats and the formats noted in Table 3 are only able to be exported from internal databases by staff.

In order for archival collections from the Royal Archives to be made accessible in a meta-aggregator such as Archives Hub or Archives Portal Europe, the data must be exported from the database rather than be made accessible through the online view of the collections. Archives Hub requires data to be provided in EAD 2002 in order to be made available through their aggregator. In order to make this data accessible to Archives Hub or any other aggregator, the Royal Archives would need to export the data directly from their database. While the Royal Collection Trust has aspirations to provide data to meta-aggregators such as Archives Hub in the future, they have not formalized a partnership with any of the meta-aggregators noted in Figure 2 at time of publication.

4. Levels of Conceptual Interoperability Model and Georgian Papers Programme

4.1 Applying the LCIM to digital libraries

In order to determine the level of interoperability that may be possible for the institutions currently and to determine what might be possible with additional work, the LCIM is expressed with examples from digital library environments. These are not meant to be a prescriptive or a complete list of possible applications but to provide select common examples.

Table 4: LCIM applied to a digital library ecosystem (Tolk et al. 2009)

Level	Layer Name	Contents clearly defined	Description of level	Examples of level in cultural heritage institutions
L6	Conceptual	Documented conceptual model	Interoperating systems at this level are completely aware of each other's information, processes, contexts, and modeling assumptions.	
L5	Dynamic	Effect of information exchanged	Interoperating systems are able to re-orient information production and consumption based on understood changes to meaning, due to changing contexts.	PERICLES project
L4	Pragmatic	Context of information exchanged	Interoperating systems will be aware of the context (system states and processes) and meaning of information being exchanged.	ontologies
L3	Semantic	Content of information exchanged	Interoperating systems are exchanging a set of terms that they can semantically parse.	common semantic model
L2	Syntactic	Format of information exchanged	Have an agreed protocol to exchange the right forms of data in the right order, but the meaning of data elements is not established.	common syntax within systems (i.e. XML, JSON)

L1	Technical	Symbols of information exchanged	Have technical connection(s) and can exchange data between systems	HTTP, FTP connection within system
L0	No	NA	NA	

Beginning at the L1, the LCIM can be explored through common methods of connecting data and systems within a digital library ecosystem. Level 0 outlines systems and methods that are not connected. Level 1 allows for systems that are connected through a shared protocol such as HTTP or FTP. The data can be copied or shared from one system to another but additional information about the data that was shared may not be understood after transfer. This is expanded in L2 by allowing for a shared formatting of data between systems that are connected via a shared protocol. By employing a common syntax, the data is better understood, although context may still be lost during transfer. By employing shared semantics in L3, the systems are better able to understand the content that has been exchanged. Common semantics expressed using the same data exchange format allows for this level of interoperability. L4 expands this to include an understanding of the context of the information that is exchanged. A common ontology can allow the systems to better understand the structure of the data. The terms are enriched through greater context allowing for shared terms to be better understood by distinct systems.

In order to achieve L5, systems would need to be able to capture information that aids in the understanding of metadata within the systems through time. While not applied to metadata specifically but to digital preservation and linked data more broadly, the PERICLES Project (2017) aimed to develop tools that could be used to document evolution in a digital ecosystem. The project uses linked open data technologies to capture information about digital objects in the environment in which they exist as well as changes to this environment and the digital objects. This example is the most closely related to L5 of the LCIM, but further exploration may elicit other examples of projects that capture changes to context in metadata through time. L6 requires a shared conceptual model that allows for systems to understand each other's processes, assumptions, and constraints. This level is still aspirational within a digital library ecosystem and does not have any examples in production. Systems for maintaining awareness of processes, contexts, and assumptions in a digital library environment are not yet available, but this level can still serve as a goal to work toward when developing systems.

While the goal of this maturity model is to measure metadata interoperability, it is not possible to measure metadata interoperability without also recognizing the potential for systems interoperability. A key requirement within Nilsson's (2010) definition of metadata interoperability is that it exists within the context of data transferred between two or more systems. So, while L6 is seemingly concerned with systems more than the underlying data, the systems and their capabilities must be required when determining the level of interoperability.

4.2 Interoperability between the Georgian Papers Programme and Library of Congress

Within the digital library ecosystem of the Georgian Papers Programme and the Library of Congress different levels within the LCIM can be achieved depending on the amount of resources that could be applied to a data transformation project. All collections can be exchanged at L1 or L2 without any additional changes made. All collections can be represented through XML and the systems are able to employ common protocols for data transfer. L3 could be achievable through the development or application of a crosswalk to a single metadata schema. As all collections have metadata expressed in XML this metadata can be transformed from many schemas into one allowing for L3 interoperability. L4 could be achievable through a shared ontology to which existing metadata could be mapped.

L4 may be the highest level which the selected collections may be able to achieve in this model, although even achieving L4 would require significant data transformations. While the Library of Congress has documented descriptive practices, the Royal Collection Trust collections

may be more difficult to transform because not all of their descriptive practices follow documented guidelines. Mapping all of these collections to a shared ontology would require a significant amount of effort to recreate the context in which the collections were originally cataloged.

5. Conclusions

The Levels of Conceptual Interoperability Model can serve as a maturity model for digital libraries that seek to improve their access to collections. Providing a leveled approach to interoperability and highlighting key improvements for each subsequent level can provide institutions with incremental changes that they may be able to complete in order to increase access. Using the examples set forth, the LCIM can be applied to current practices in digital libraries and can be used to measure the potential for interoperability between collections.

While overall a useful model for understanding and marking interoperability between digital library systems, the application of the LCIM may be limited because of the confusion that could arise from terms used to describe it. L3 is labeled "Semantic Interoperability" yet the description of the level is less complex than the understanding of semantic interoperability that information professionals may have already. In fact, semantic interoperability, a concept that stems from the development of the Semantic Web, is more similar to what the LCIM labels as Conceptual Interoperability or L6 of the model.

This analysis focused specifically on metadata interoperability although that is only one part of a digital library ecosystem. With the development of the International Image Interoperability Framework (IIIF), image interoperability in a digital library setting is a potential place of further exploration for this analysis. Additionally, further exploration of interoperability between specific systems that are commonly used in cultural heritage materials would provide a wider view of the potential for interoperability within a digital cultural heritage ecosystem.

An important aspect of this approach to interoperability research is that collection materials should be analyzed through the framework of a model that does not limit the potential for interoperability. Level Six of this model should not be thought of as the upper limit for composability or interoperability. As technology continues to develop and allow for further markers of interoperability, the LCIM should be expanded to include subsequent levels.

References

Acting on Change: Model-driven Management of Evolving Digital Ecosystems (Rep.). (2017). PERICLES. doi:http://pericles-project.eu/uploads/files/PERICLES_White_Paper_Acting_On_Change_2017.pdf

Alipour-Hafezi, M., Horri, A., Shiri, A., & Ghaebi, A. (2010). Interoperability models in digital libraries: an overview. *The Electronic Library, 28*(3).

Chan, L. M., & Zang, M. L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I. *D-Lib Magazine, 12*(6). Retrieved from http://dlib.org/dlib/june06/chan/06chan.html

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. (2002). Metadata Principles and Practicalities. *D-Lib Magazine, 8*(4). Retrieved from http://www.dlib.org/dlib/april02/weibel/04weibel.html

Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys, 42*(2). Retrieved from http://eprints.cs.univie.ac.at/79/1/haslhofer08_acmSur_final.pdf

Heery, R. (2004). Metadata Futures: Steps toward Semantic Interoperability. In D. Hillman & E. Westbrooks (Eds.), *Metadata in Practice* (pp. 257-271). Chicago, IL: American Library Association.

Nilsson, M. (2010). *From interoperability to harmonization in metadata standardization designing an evolvable framework for metadata harmonization* (Unpublished doctoral dissertation). Diss. (sammanfattning) Stockholm: Kungliga Tekniska högskolan.

Nilsson, M., Baker, T., & Johnston, P. (2009). *Interoperability Levels for Dublin Core Metadata* (Publication). Dublin Core Metadata Initiative. doi:http://dublincore.org/documents/interoperability-levels/

Royal Collection Trust (Ed.). (2017). Georgian Papers Online. Retrieved from http://gpp.royalcollection.org.uk/What.aspx

Tennant, R. (2001, February 15). Different Paths to Interoperability. *Library Journal*, *126*(3).

Tolk, A., Wang, W., & Wang, W. (2009). The Levels of Conceptual Interoperability Model: Applying Systems Engineering Principles to M&S. *Spring Simulation Multiconference.*

Van de Sompel, H., & Nelson, M. L. (2015). Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine, 21*(11/12). Retrieved from http://dlib.org/dlib/november15/vandesompel/11vandesompel.html

Washington, G. (n.d.). Unpublished manuscript, George Washington Papers.

Zeng, M., & Qin, J. (2016). *Metadata. 2nd, rev. ed.* Chicago, IL: American Library Association

A Data Model for Lifecycle Management of Natural Hazards Engineering Data

Presentation

Maria Esteva
University of Texas at Austin, USA
maria@tacc.utexas.edu

Ashley Adair
University of Texas at Austin, USA
a.adair@austin.utexas.edu

Craig Jansen
University of Texas at Austin, USA
cjansen@tacc.utexas.edu

Josue Balandrano Coronel
University of Texas at Austin, USA
jcoronel@tacc.utexas.edu

Sivakumar Ayeegoundanpalay Kulasekaran, University of Texas at Austin, USA
siva@tacc.utexas.edu

Keywords: research data lifecycle; Fedora 4 repository; multi-structured metadata

Abstract

Natural Hazards engineering data derives from sophisticated experimental design and contains a complex array of relationships. Representing and publishing these data is challenging, as the domain lacks a metadata schema and specialized vocabulary. To build the functionalities required to curate and publish datasets within the DesignSafe-CI, an end-to-end research data lifecycle platform (https://www.designsafe-ci.org), the curation team took a multi-step approach.

First, the team undertook modeling of the research processes of seven kinds of experimental projects and corresponding hazard types that are supported by the CI. Researchers in the space were asked to draw and describe their research workflows, noting the equipment, the processes involved and their output data, the software used to analyze the data, and the documentation that are indispensable for proper data interpretation and reuse. To derive a generic experimental data model, the team analyzed these workflows and identified common processes as well as the relationships between those. The activity arrived at core metadata elements that represent the steps and methods involved in Natural Hazards projects, as well as sets of user-suggested vocabularies specific per experiment type. The resultant data model emphasizes the datasets structure and the provenance of the multiple data outputs obtained from different configurations within an experimental project. Definitions for the core metadata and vocabularies are maintained in the community meta-dictionary YAMZ (www.yamz.net).

In the DesignSafe-CI portal the data model was implemented as interactive functions that allow users to progressively tell the story of their project by categorizing, describing, and relating data from an experiment. Using the metadata and the vocabularies users can start and stop working with their data at any time: selecting and deselecting files, adding or removing categories, and editing descriptions. As users go about curating their files in the interface, the network of relations of the experiments is formed in the back-end through a middleware metadata API. This allows rendering the experiments graphically as a tree showing the links between processes, data, and descriptive tags and narratives; helping users arrive to the decision of publishing.

At the point of publication, the final data and metadata transitions to a Fedora 4 repository. For this, we mapped each of the elements and terms from the data model to three metadata schemes: Dublin Core, to describe the experimental project; PROV to represent provenancial relationships among processes and their outputs; and DataCite for metadata that will be passed in the minting of Digital Object Identifiers (DOIs). Mapping across these schemes results in multi-structured metadata that standardizes elements and vocabularies. Beyond description and contextual

information as minimum requirements to publish scientific data, this data model emphasizes the structure of the experiments and uses terms familiar to users in the domain to facilitate data reuse. Its mapping to standard schemas enable proper publication, exchange, and web exposure of the data, and allows queries that relate the components. Friendly user evaluations conducted for the preliminary release of the curation and publishing pipelines suggest that they are intuitive and will be complemented with a larger study in the Fall.

Best Practices for Software Metadata: A Report from the Software Preservation Network

Presentation

Elizabeth Roke
Emory University, U.S.A.
elizabeth.roke@emory.edu

Daniel Noonan
Ohio State University, U.S.A.
noonan37@osu.edu

Keywords: Software Preservation Network; survey of metadata practices; crosswalk of software metadata standards

Abstract

Representatives of the Software Preservation Network Metadata Working Group will present preliminary results from its analysis of metadata for software preservation, discussing a survey of existing practices as well as a crosswalk of existing ontologies and schemas in use in libraries, the open source software community, and specialized research communities. The talk will also discuss the activities of the Software Preservation Network, highlighting activities that have a direct impact on metadata practice.

Discussion

The Software Preservation Network (SPN) is an effort to coordinate software preservation efforts to ensure the long-term access to software. This work currently involves legal licensing and information policy research; an international registry of software collections; and software development contributions to technical infrastructures that facilitate long-term access to software. The SPN Metadata Working Group is focused on developing, promoting, and advocating for common metadata frameworks and related metadata standards, vocabularies, and ontologies that support software discovery, preservation, and access.

Currently, there is a wide range of metadata software practice, from descriptive metadata encoded in library-based MARC records to user-created metadata for open source software packages documented in software repositories such as GitHub. Although there have been a few projects that recognize the need to improve metadata standards for software, particularly in the data community, no common framework for thinking about software metadata has emerged that addresses all required metadata semantic elements. This is especially important when we consider the different use cases for software metadata throughout its lifecycle, from creation to preservation and emulation.

This presentation will report on two recent initiatives of the SPN Metadata Working Group: an international survey of metadata practices regarding software and a crosswalk designed to map different software metadata standards together in a single framework. We will summarize the existing metadata landscape for software metadata, including approaches within the library, data research, and open source community, and explore the metadata lifecycle for software with special attention to its unique preservation needs and challenges.

Teaching and Learning

LD4PE: A Competency-based Guide to Linked Data Principles and Practices

Stuart A. Sutton
Dublin Core Metadata Initiative, USA
sasutton@dublincore.net

Michael D. Crandall
University of Washington, USA
mikecran@uw.edu

Marcia Zeng
Kent State University, USA
mzeng@kent.edu

Thomas Baker
Dublin Core Metadata Initiative, Germany
tom@tombaker.org

Abigail Evans
University of Washington, USA
abievans@uw.edu

Sean Dolan
Kent State University, USA
sdolan5@kent.edu

Joseph Chapman
D2L, Ltd., USA
Joseph.chapman@D2L.com

David Talley
Precise Recall, USA
dtalley@preciserecall.com

Michael Lauruhn
Elsevier Labs, USA
M.Lauruhn@elsevier.com

Abstract

The IMLS-funded project *Linked Data for Professional Education* (LD4PE) has prototyped a competency-based referatory of Learning Resources related to the design, implementation, and management of Linked Data (a referatory is a website that points to and describes learning resources.) The project developed: 1) A "Competency Index for Linked Data" (*Index*), which characterizes the Linked Data field in terms of formally identified competencies usable for tagging resources, as a basis for self-guided learning or for designing curricula. 2) A tool set for creating metadata about Learning Resources, packaging user-selected Learning Resources in Saved Sets, and creating learning trajectory maps for expressing personal or curriculum-based learning journeys through the competencies. 3) A catalog of Learning Resources mapped to the competencies. 4) A project website, http://explore.dublincore.net. 5) Best Practices describing policies reusable by other projects using competency-based description and discovery of learning resources.

Keywords: Linked Data, Learning Resources, competency frameworks, ASN-DL, competency-based teaching and learning.

1. Introduction

Understanding Linked Data standards and practices has become a key requirement for information professionals in galleries, libraries, archives and museums (GLAM). Major national libraries and bibliographic services are leading the trends toward publishing authority files, catalogs, datasets, and bibliographic standards as Linked Data, and toward aggregating Linked Data from external sources such as dbpedia and MusicBrainz to enhance discovery and retrieval. Broad initiatives such as LODLAM and OpenGLAM promote the integration of Linked Data across galleries, libraries, archives, and museums.

Cultural memory institutions find themselves on shaky ground as this paradigm shift pushes the need for competent professionals from national centers toward local institutions. The challenge of acquiring new competencies extends to teachers of the next generation of professionals and trainers who provide continuing professional development.

This urgent need to develop Linked Data competencies in the professional workforce is driving major initiatives to provide Learning Resources about the underlying standards and model of Linked Data such as the EU's Euclid project, LOD2, School for Data, PlanetData, Open Data

Institute, Lean Semantic Web, Cloudera, GATE, *Linked Data Cookbook*, and *Cookbook for Translating Data Models to RDF Schemas*. The products of these initiatives range from curricular structures and full courses to simple "recipes"—brief packages of "how-to" videos and step-by-step instructions that address specific, but frequently unarticulated learning outcomes. These scattered initiatives and their resources can be easy to find by those who already know what they are looking for, but everyone else struggles to put the available resources into a context.

The website of the LD4PE project helps alleviate this struggle by supporting the structured discovery of online Learning Resources that have been made available by both open and commercial providers. At its heart is a competency framework for Linked Data practice that supports the tagging of Learning Resources according to the specific competencies they address. The competency framework, itself expressed in RDF, leverages Linked Data technology by assigning global identifiers (URIs) to statements of competence. These URIs, when used as tags in metadata about Learning Resources, map the Learning Resources to nodes in the *Index*.

Deliverables of the LD4PE project include:

1. *Competency Framework*. RDF-modeled "Competency Index for Linked Data" (*Index*) based on the Achievements Standards Network Description Language (ASN-DL) for describing formally promulgated competencies and benchmarks.
2. *Toolkit*. An openly available, Web-based tool set to support the management of the *Index*; the generation of RDF metadata about Learning Resources; the packaging and arrangement of selected Learning Resources by users in Saved Sets; and the creation of learning trajectory maps expressing curricular structures or personal learning journeys superimposed over the competency framework through the integration of these elements as WordPress custom posts and taxonomies on the LD4PE website.
3. *Learning Resource Descriptions*. Metadata about Learning Resources mapped to the competencies and benchmarks of the *Index* in support of competency-based resource discovery by teachers, trainers and learners.
4. *LD4PE Website.* A website (http://explore.dublincore.net), ownership of which is being transferred to the Dublin Core Metadata Initiative (DCMI) as part of its educational agenda.
5. *Best Practices*. Readily accessible best practice documentation for all LD4PE development and maintenance processes, from the creation of a community-based competency framework development to the creation of metadata about Learning Resources and learner trajectories.

2. Competency Index

Competency frameworks defining what a learner should know and be able to demonstrate underpin labor market credentialing mechanisms including college and university degrees, educational certificates, industrial certification, occupational licenses and even emerging micro-credentialing through digital badging. Learning environments strive for a tight coupling between these controlling frameworks, the resources necessary to achieve the learning objectives embodied in the frameworks, and what is actually assessed in terms of learner competence (Sutton & Golder, 2008).

2.1. *Index* Development

Government and quasi-governmental agencies, commercial entities and professional organizations develop most such competency frameworks under the tight editorial control of selected experts. In contrast to such highly constrained contexts, the LD4PE approach aims at crowd-sourcing expertise both to develop and to assess the project's competency framework, with synthesis and finalization of the published framework in the hands of a small, volunteer Editorial Board comprised of subject matter experts in the domain of Linked Data. During the project period, members of the project team personally solicited feedback at a number of conferences, workshops and other events, and widely disseminated a call for comments on a public version of the *Index*.

Projects such as schema.org are demonstrating that useful outcomes can be achieved through forms of crowdsourcing when version control systems created for "social coding", such as Github, support processes for community input; this approach is discussed further in Section 2.5.

Figure 1 illustrates the various components in the LD4PE competency-creation process.

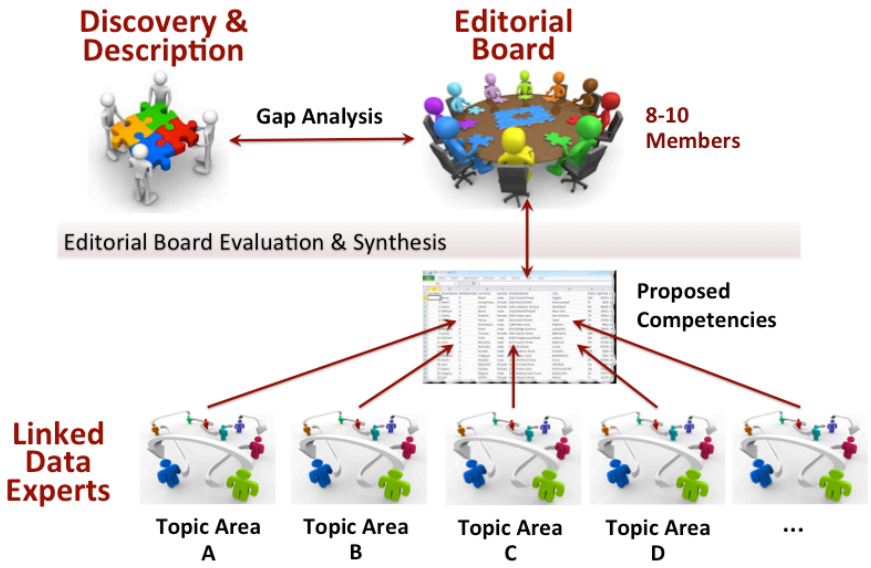


FIG. 1. Crowd-sourcing expertise for the *Index*.

Discovery and Description. The process of discovering, describing, and mapping Learning Resources to competencies involves continuous assessment of the coverage of the *Index* in terms of:

1. *orphan competencies*—competencies found in Learning Resources that cannot be mapped to the *Index*; and
2. *misaligned granularity*—where imprecise mapping options reveal a mismatch in granularity between *Index* competencies and Learning Resources.

Linked Data Expert engagement through both solicited and unsolicited proposals for new competencies provides a mechanism through a version-controlled submission process for subsequent review and synthesis into the evolving *Index* by the *Editorial Board*, a small group of domain experts who volunteer their time.

2.2 *Index* modeling

The *Index* itself is modeled as an RDF graph using the Achievements Standards Network Description Language (ASN-DL). The ASN-DL is comprised of two resource classes, *StandardDocument* for competency frameworks described as wholes, and *Statement* for individual competency assertions in the document. Object properties are used both to replicate the original hierarchical or graph modeling of the canonical document and to express the semantic relationship between individual assertions.

The ASN-DL is extensible through addition of new properties and subproperty refinements to support domain-specific profiles. Value vocabulary namespaces relevant to the jurisdiction of the competency framework may be specified and encoded as RDF using the W3C standard Simple Knowledge Organization System (SKOS).

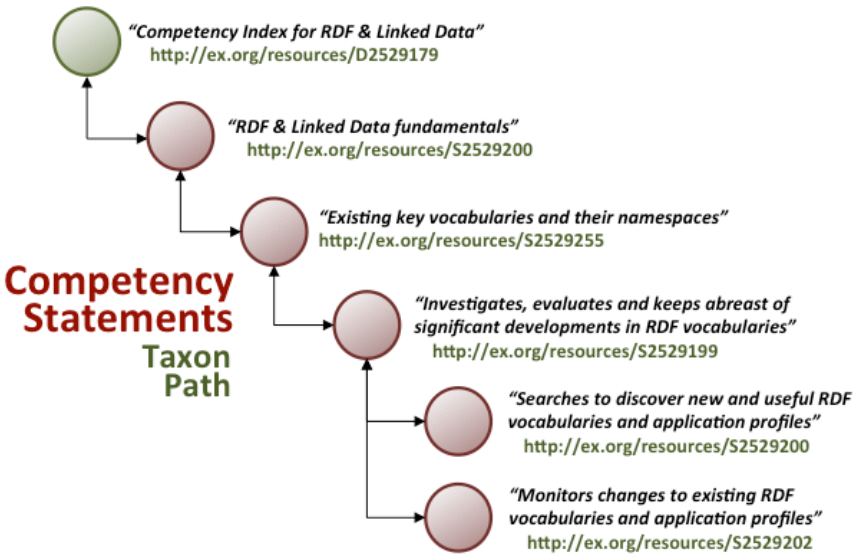


FIG. 2. Example Index ASN-DF modeling

2.3 *Index* scope

One of the major challenges during the early months of the project was operationalizing the term "Linked Data" in such a way that the scope of both the developing *Index* and the Learning Resource to be described and mapped were tractable within the context of the two-year grant. "Linked Data" denotes a method for publishing structured data on the web that builds on a set of open standards and technologies for expressing and querying of Linked Data, such as HTTP, URI, RDF, SPARQL, OWL, and SKOS. The concept of Linked Data is roughly characterized in terms of four practices recommended by Sir Tim Berners-Lee: the use of URIs as names for things; the use of HTTP URIs that allow people to find information about those names; the provision of that information in a machine-readable form; and the inclusion of links to other relevant resources (Berners-Lee, 2006). Extending the *Index* and the body of Learning Resources described in LD4PE metadata to encompass all underlying and enabling technologies for Linked Data would have been unrealistic as the scope for a single competency framework. The project struggled with the issue of where to draw the line. The community aspect of the project is meant to provide a continuous feedback channel on this issue and foster an evolving sense of the boundaries of the *Index.*

2.4 Outcome

The outcome of this process is a community-developed *Index* that describes a set of learning objectives and outcomes in terms of relevant knowledge, skills, practices, and habits of mind necessary to learn successful Linked Data practice, from design and modeling through implementation and maintenance.

The structure of the CI is as follows:

- *Topic cluster*
  - **Topic**
    - Competency: Tweet-length assertion of knowledge, skill, or habit of mind
      - *Benchmark: Action demonstrating accomplishment in related competencies*

This structure is implemented in the LD4PE triplestore, and expressed on the Learning Resources page of the LD4PE website. There are six Topic Clusters in the current version:

1. Fundamentals of Resource Description Framework
2. Fundamentals of Linked Data

3. RDF vocabularies and application profiles
4. Creating and transforming RDF Data
5. Interacting with RDF Data
6. Creating Linked Data applications

Under these clusters, there are 30 topic groups, which contain 95 competencies. In the following example, under the topic "RDF serialization", there are two competencies, each of which is associated with a benchmark.

- **RDF serialization**
  - Distinguishes the RDF abstract data model and concrete serializations of RDF data.
    - *Expresses data in serializations such as RDF/XML, N-Triples, Turtle, N3, Trig, JSON-LD, and RDFa.*
  - Understands RDF serializations as interchangeable encodings of a given set of triples (RDF graph).
    - *Uses tools to convert RDF data between different serializations.*

One can think of competencies as expressing the knowledge to be imparted by instructors, covered in tutorials, or acquired by self-learners, and benchmarks as the basis for homework assignments, exercises, or exams.

2.5 Further Development of the Competency Index

Like the evolving technologies of Linked Data itself, the Competency Index for Linked Data will always be a work in progress. Over the course of the project, technologies for supporting a crowdsourced approach to the maintenance of the Index have also evolved. As of September 2017, the "canonical" version of the index has been installed in a Github repository under management of DCMI (in both a Chinese and English version). Github is a "social coding" platform originally created to support the collaboration of programmers in the development of open-source software such as Linux, but is increasingly being used for collaboration on other open-source endeavors, such as the maintenance of metadata vocabularies such as Schema.org.

The LD4PE Competency Index Editorial Board will continue its work in this new context, using Github's facilities for allowing collaborating editors, or even members of the public, to propose changes or additions to the Index. By navigating to https://dcmi.github.io/ldci/D2695955/, anyone can click on the button "Edit in Github", after which they will be guided through the process of submitting a "pull request" -- a request that the changes be integrated into the master copy of the Index. Github will display the differences between the master copy and the copy with proposed changes, side-by-side, for consideration by the Editorial Board; support discussion threads about aspects of the proposal; and accepted changes will be merged into the master copy.

LD4PE and DCMI have not yet established any policy about the URIs used to identify competencies or to their persistence, so the ASN-based URIs hitherto used may be replaced by new URIs, such as PURLs. DCMI also wants to experiment with new approaches to keeping translations of the Index synchronized as changes are made. The Chinese translation has been set up its own Github repository and for now, changes to the English version will be communicated to the Chinese translators in the form of "diffs" -- side-by-side comparisons showing changes approved and published by the Editorial Board.

Although no formal assessments have been made of the Competency Index to date, the development process itself has provided substantial input from subject matter experts as well as users. We expect that further use and adoption of the Index will contribute to its maturation and development over time.

3. Toolkit

The LD4PE project is committed to the use of open standards and, where appropriate, adaptation of existing tools and toolsets, implementing anew only where necessary. The LD4PE toolkit is a set of open tools that enable the expression and use of the *Index* and associated Learning Resources through the LD4PE website. Key areas of project implementation are light-weight, browser-based editors for creation of Learning Resource descriptions, competency descriptions, and controlled vocabularies. On the LD4PE website, custom WordPress plug-ins manage integration of Learning Resource descriptions as instances of custom post-type and represent competency descriptions and vocabularies as custom taxonomies. Custom plug-ins also extend WordPress functionality to allow authenticated users to save collections of Learning Resource descriptions and organize competency statements into "learning maps" sequenced to support curriculum development.

3.1 Metadata Generation Tools

To support the generation of metadata describing Learning Resources and for authoring the *Index* and its competency assertions, the project has developed an open toolkit consisting of two browser-based editors implemented using AngularJS. The first of these editors is designed for creating RDF Learning Resource descriptions based on an application profile of the Learning Resource DCMI (schema.org) schema.

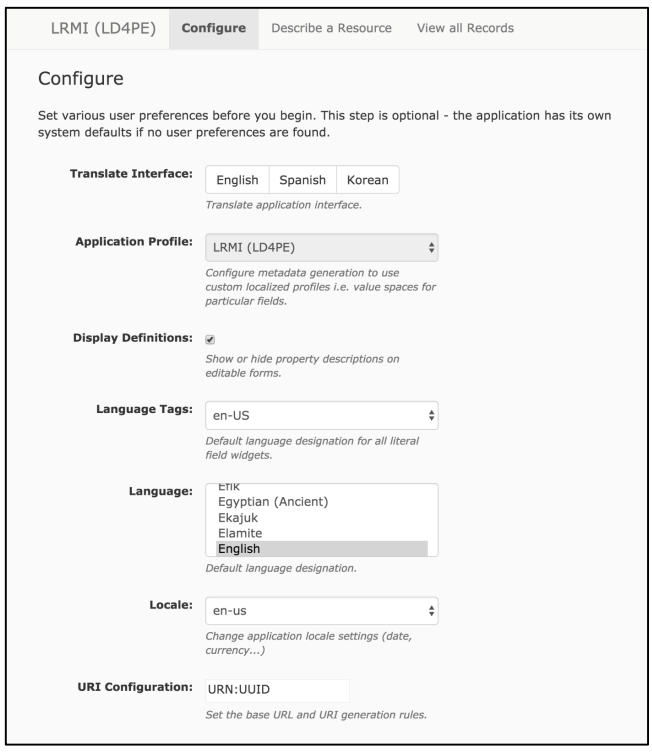


FIG. 3. LD4PE Competency Framework editor.

The second editor can be used to create *Index* metadata based on the ASN Description Language (ASN-DL) for competency framework modeling and description, following the editorial process described earlier, or some other version of that process suitable to another knowledge domain.

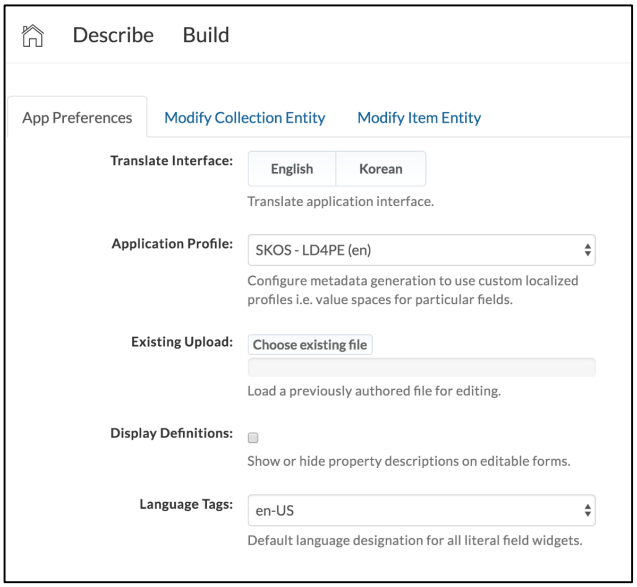


FIG. 4. LD4PE Learning Resource editor.

These two editors are key components of the reusable design followed by the project. While the LD4PE project has focused exclusively on the domain of Linked Data, we anticipate that others may wish to create similar competency frameworks for other areas, and catalog Learning Resources related to those domains using these tools. The tools and custom WordPress plug-ins will be available for download from Github to allow their reuse in other contexts

3.2 Saved Sets and Learning Maps Tools

While the *Index* defines a set of competencies, it neither prescribes any competencies as "core" nor defines a logical sequencing of its components. In other words, the *Index* does not, in itself, define a curriculum by prescribing a specific learning trajectory or map through the set of competencies. Instead, the LD4PE project provides tools to enable teachers, trainers, and learners to map their own pathways through the *Index* graph, through creation of Saved Sets of Learning Resources and *Learning Maps* of competencies. Saved Sets and Learning Maps traverse, or overlay, the competency nodes of the *Index*.

This is accomplished through custom plug-ins that extend basic WordPress functionality to allow authenticated users to define *Saved Sets* of Learning Resource descriptions for future access, as well as adding and removing Learning Resources from their sets. A similar tool allows users to sequence sets of competency statements to define Learning Maps representing logical pathways for instruction. Both tools give users the option to open sets or maps for public access, allowing other users to take those curated collections as starting points to extend and adapt for their own purposes.

The saved learning trajectories or pathways may be identified as formal curriculum structures or as personalized trajectories created by instructors or learners as evidence of progress. They may also function as suggested paths forward in the learning process or as guides to other instructors or learners in creating their own trajectories.

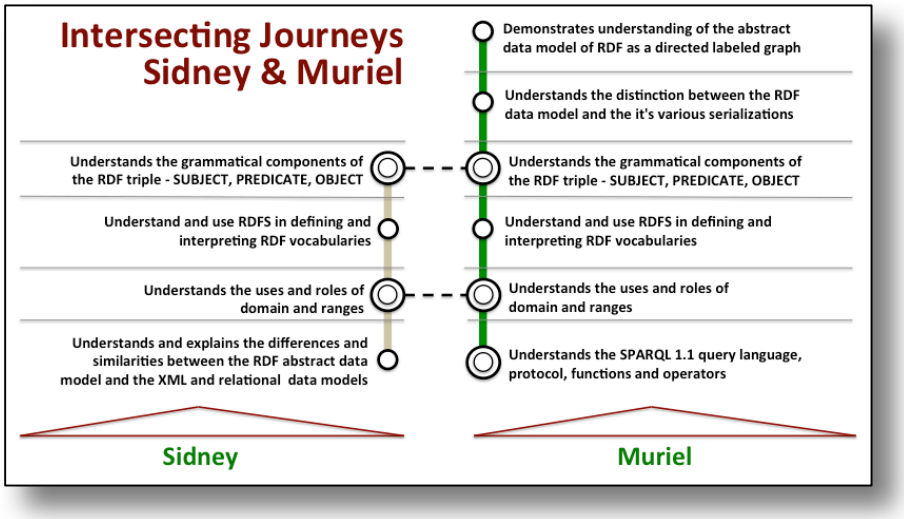


FIG. 5. Learning trajectory maps.

Research has shown that such maps reveal the macrostructure of the body of information or knowledge within a field, making the context of that information or knowledge more apparent and useful to learners (O'Donnell et al., 2002; Hall & O'Donnell, 1996; Hall & O'Donnell, 2010). Similarly, learning trajectory overlays contextualize competencies and learning outcomes in a "larger picture" of learning expectations, outcomes, and personal learner progressions.

Thus, LD4PE stands in sharp contrast to other projects in not prescribing a single curricular point of view but in providing instead the means for instructors, trainers and learners to chart multiple, diverse pathways for learning—pathways defined as public or private, individual or collective, prescribed or exploratory. Each Saved Set provides a different roadmap for discovering and traversing lesson plans, how-to recipes, webinars, and tutorials that have been described and aligned to the competency nodes of the *Index* graph.

3.3 Sample data set for teaching

Although not formally part of the tool kit, OCLC has contributed a large static triplestore (a subset of their WorldCat Linked Data focused on the Library Science domain) that is available for use in the creation of stable examples and assessments against the competencies that can be used in teaching situations. This avoids the difficulties inherent in using live examples which can change rapidly from minute to minute, and allows instructors to create reusable activities for learners with known outcomes. A tutorial has also been created to help novice users take advantage of this data set and construct queries against the triple store (http://explore.dublincore.net/related/oclc-dataset/).

4. Architecture

The LD4PE technical architecture consists of a WordPress public interface on top of a triplestore managing Learning Resource descriptions, the *Index* itself and controlled vocabularies for selected statements. As of September 2017, the *Index,* hitherto available through the ASN-D2L open repository and reflected on the LD4PE WordPress website as a custom taxonomy, is henceforth available at https://dcmi.github.io/ldci/, backed by the DCMI Github repository https://github.com/dcmi/ldci/. The editing tools function as open, stand-alone applications that can be used independently of the WordPress instance and the project's triplestore.

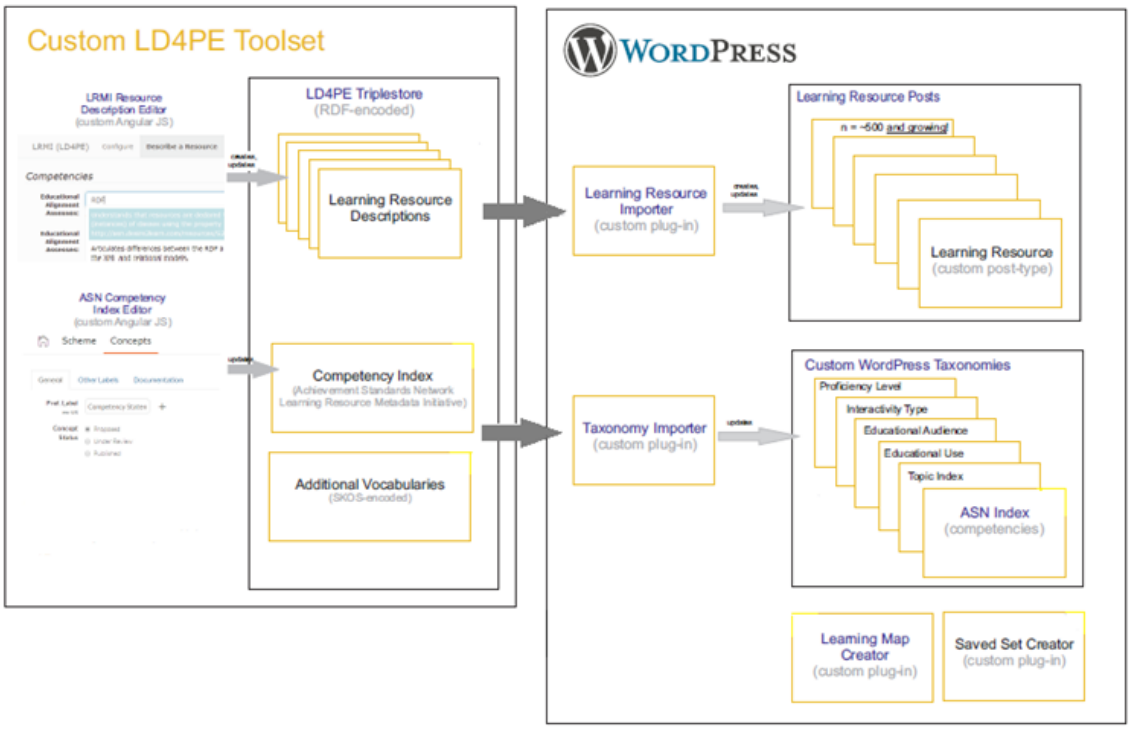


FIG. 6. LD4PE architecture.

The LD4PE website primarily provides competency-based browse access to Learning Resources. In addition, it provides a publication venue both for learning trajectory maps and for select Learning Resources such as a pilot set created by project partners Sungkyunkwan University Institute of Information and Management (Korea), OCLC, Elsevier, Synaptica, and Access Innovations. A key asset donated by OCLC is a very large static triplestore publicly available as a resource for developing replicable exercises, tests, and examples for teaching Linked Data principles.

5. Best Practice Documentation and Dissemination Activities

Publicly available guidelines are published on the LD4PE website covering best practices in:

1. Community development and management of competency frameworks, including a description of the process used by the editorial board to define and organize the competencies and benchmarks, and achieve a consistent style across the range of entries. (English version [https://dcmi.github.io/ldci/]; Chinese version [https://dcmi.github.io/ldci-zh/]);
2. Creation of useful metadata descriptions of Learning Resources using the Learning Resource DCMI application profile and alignment of Learning Resources to relevant competencies, including instructions for using the publicly available schema tools (http://explore.dublincore.net/related/share-our-tools/);
3. Development of learning pathways to create guided navigation through the competencies by teachers and learners, for use in curricula or training sessions aimed at various audiences and knowledge levels (http://explore.dublincore.net/theory/learning-pathways-as-transit-maps/); and
4. A sample data set that can be used for teaching purposes, along with a tutorial that can be used to help novice users take advantage of this data set and construct queries against the triple store (http://explore.dublincore.net/related/oclc-dataset/).

Through international professional conferences, the LD4PE team members have reached out to various communities that may benefit from the Competency Index and the Learning Resources compiled by the LD4PE. These include the schools of information science (iSchools), the galleries, libraries, archives and museums communities, and scholarly publishing community. Conference presentations and training sessions have been given in the United States, Europe, and Asia. In addition, extensions of the Learning Resources and *Index* to other languages has been encouraged and supported, as demonstrated by the first fully translated Competency Index to Chinese in 2016, rolled out in a well-attended seminar in Asia. The tool set of LD4PE supports the configuration of non-English languages for the resource description editor as well, providing the opportunity for inclusion of Learning Resources in non-English languages in the future.

6. Conclusion

While LD4PE focuses specifically on skills, knowledge, and professional practice in the area of Linked Data, nothing precludes the extension of its competency-based approach to other areas of library and information science (LIS), archives, and museum curricula. Similar competency frameworks could be developed describing practice in knowledge organization systems, cataloging, and organizational management. Describing a knowledge domain in terms of competencies necessary for mastery provides a solid scaffolding for teaching and learning, and our hope is that others will take advantage of the tools and practices embodied in this project to develop similar offerings in other areas.

This approach will be appropriate wherever the goal is for learners to achieve competence within a defined set of knowledge, skills, and habits of mind. DCMI is exploring ways to sustain and grow the products and assets of this work moving forward because it sees the project's conceptualization and its outcomes (including best practice documentation related to the framework itself) as generalizable to all areas of principled metadata design and best practice, while also providing a guiding template for future development of DCMI's education and training agenda.

Acknowledgments

The work described in this poster was partially funded by the Institute of Museum and Library Services (IMLS). Content partners Sungkyunkwan University Institute of Information and Management (Korea), OCLC, Elsevier, Synaptica, and Access Innovations were generous in donating their time and expertise to the project.

References

Berners-Lee, Tim (2006). Design Issues: Linked Data. Retrieved June 28, 2015 from http://www.w3.org/DesignIssues/LinkedData.html.

Crandall, M.D., Tennis, J., Sutton, S.A., Baker, T. & Talley, D. (2013). Planning a platform for learning linked data. In Moen, W. & Rushing, A. (Eds.), *International Conference on Dublin Core and Metadata Applications*. Retrieved June 28, 2015 from http://dcpapers.dublincore.org/pubs/article/view/3693/1916.

Hall, R.H. & O'Donnell, A. (1996). Cognitive and affective outcomes of learning from knowledge maps. *Contemp. Educ. Psychol*. *21*, 94-101.

Hall, R.H., Hall, M.A. & Saling, C.B. (2010). The Effects of Graphical Postorganization Strategies on Learning from Knowledge Maps. J. Exp. Educ. 67(2):101-112 (DOI:10.1080/00220979909598347).

O'Donnell, A.M., Dansereau, D.F. & Hall, R.H. (2002). Knowledge maps as scaffolds for cognitive processing. *Educ. Psychol. Review 14*, 71-85.

Sutton, S. A. & Golder, D. (2008). Achievement Standards Network (ASN): An application profile for mapping K-12 educational resources to achievement standards. In Heike Neuroth (Ed.), *Proceedings of the International Conference on Dublin Core and Metadata Applications*, (pp. 69-79). Retrieved, June 28, 2015 from http://dcpapers.dublincore.org/ojs/pubs/article/view/920/916.

Ward, N. & Nicholas, N. (2010). Benefits of machine readable curricula. Retrieved, June 28, 2015 from http://www.australiancurriculum.edu.au/static/docs/Benefits%20of%20a%20Machine%20Readable%20Curricula.pdf

Understanding Users' Metadata Needs: How Do We Know What They Want?

Presentation

Jeanette Norris
Brown University Library,
USA
jeanette_norris@brown.edu

Keywords: library metadata; understanding user expectations; MARC

Abstract

Descriptive metadata should match users' expectations of the information that is available to search against. Some standard approaches to understanding user needs related to library metadata include usability tests on existing search interfaces, mining search logs and other similar sources for information about the search strategies that people use. Those strategies rarely separate the raw data from how the system uses it and they provide little data that contextualizes the users' choices of search terms. The methodology used for the research discussed in this presentation focuses on how users describe books outside of the context of an existing search interface. It represents an effort to isolate and identify salient types of information and then to compare them with library data and standards to determine how much users' descriptions and catalogers' descriptions overlap.

In the summer of 2016, the presenter ran pilot study to explore what types of information users' find most important for being able to find, identify, and select books. The six university students who participated were asked to write free-text descriptions of three books that would allow someone else to be able to find, identify, and decide whether or not they wanted to read the books. The free-text descriptions that participants wrote during the sessions were compared with MARC records that had been created for the same three books and were in the library's online catalog. The descriptions created by participants can be used to glimpse into the needs of users as they approach online search interfaces trying to find a resource of interest. The study focuses on the descriptive metadata itself, rather than on the effectiveness of the functionality created with that metadata. Additionally, by removing the search interface we can analyze two relatively similar descriptions that also provide comparable levels of contextual information.

This presentation will focus on an analysis of the methodology used in this and other similarly constructed studies, preliminary findings based on the data that was gathered during the pilot study, and ideas for how the type of information gathered through these types of studies could be used to assess metadata practices and inform the creation of descriptive metadata standards.

Semantic Web Workbench:
Tools, Ontologies, Software

Metadata for Improving Transparency in the Credentialing Marketplace

Presentation

Jeanne Kitchens
Southern Illinois University, USA
jeannekitchens@siu.edu

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

Robert G. Sheets
George Washington University, USA
bobgsheets@gmail.com

Keywords: credentials, metadata validation, Resource Description Framework (RDF), Linked Data, workforce development

Abstract

Summary

The Credential Transparency Description Language (CTDL) is a family of Resource Description Framework (RDF) specifications for describing and relating resources in the Web's credentialing ecosystem. The goal of the development is to bring transparency to a chaotic, ill-defined, high-stakes environment of credentials by means of rich metadata descriptions and equally rich linking of resources. The CTDL definition of "credential" is broad and includes subclasses ranging from certifications through formal college and university degrees to micro-credentials and badges. In addition to credentials, the family of specifications includes RDF schemas for credential-related classes such as learning opportunities, assessments and competencies. The family of specifications builds on schema.org types (hereafter classes), their associated properties, and schema.org's intentionally weak semantics. While the focus of the CTDL is solely on the description of credentials and related entities as abstract works; it is anticipated that CTDL descriptions may be used by others to definitively identify and verify credential instances awarded to individuals.

Discussion

Credential Engine[1] is a U.S. 501(c)(3) non-profit organization whose mission is to improve transparency in the credentialing marketplace by making it possible to discover and compare credentials in an exploding across the credentialing ecosystem. In the *Credential Engine* context, credentials are broadly defined on a continuum from university degrees, through certificates and certifications to digital badges. *Credential Engine* pursues its mission by promoting an open applications marketplace through maintenance of the open-licensed Credential Registry (CR) and the Credential Transparency Description Language (CTDL). *Credential Registry* uses both the open metadata infrastructure of the CTDL and open-licensed software to continuously capture, connect, archive and share metadata about credentials, credentialing organizations, quality assurance organizations, and competency frameworks, as well additional metadata as needed to support an open applications marketplace.

Credential Engine grew out of the Credential Transparency Initiative (CTI), which began in 2013. Supported by the Lumina Foundation, CTI was led by the George Washington University's Institute of Public Policy (GWIPP), Workcred—an affiliate of the American National Standards Institute (ANSI), and Southern Illinois University (SIU) Carbondale's Center for Workforce Development. CTI worked closely with an Executive Committee, a Technical Advisory Committee, and hundreds of diverse credentialing stakeholders that provided feedback focused on three components: (1) development of the CTDL, (2) development of a prototype Credential

[1] https://www.credentialengine.org/

Registry, and (3) development of a prototype interface to the Registry called Workit™ that is scheduled for launch in December 2017.

The primary classes and their general relationships are illustrated in Figure 1.

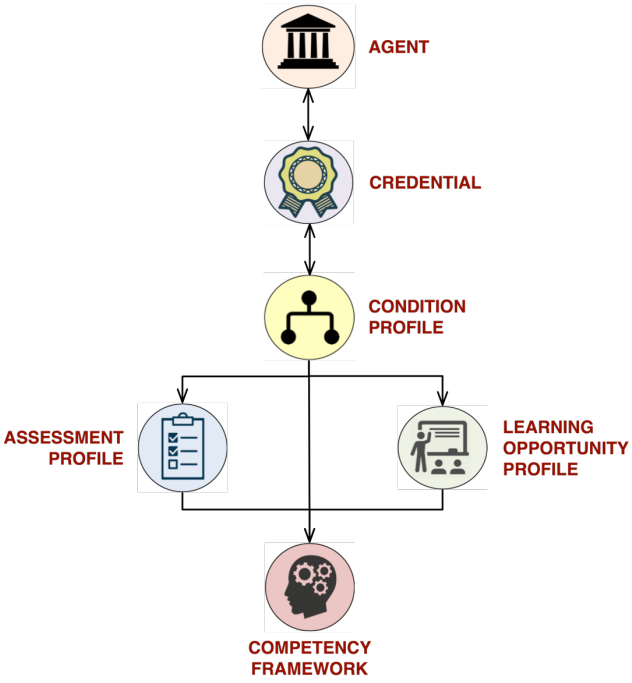


FIG 1. Primary CTDL classes

The CTDL family of RDF specifications is comprised of approximately 50 classes, 236 properties, 18 concept schemes and 128 concepts. The credential class, its array of subclasses, and its associated learning opportunities, assessments and supplemental profiles are expressed using the CTDL description language.[2] Competencies are expressed using a *Credential Engine* profile[3] of the Achievement Standards Network description language.[4] Currently, metadata describing credentials and related classes are serialized and available in JSON-LD and Turtle.

The management of the RDF specifications is handled using a Neo4J application that, from a single console: (1) manages updates to the schemas; (2) generates all of the human readable schema documentation pages with the exception of the narrative user guides; (3) generates the schema serializations; and (4) generates the JSON schemas validating credential metadata destined for the Credential Registry. In the first quarter of 2018, the application will be updated to handle validation through ShEx[5] and SHACL[6] expressions.

The CTDL was first released on November 17, 2016 through a collaborative process with a large Technical Advisory Committee and based on information and feedback from over 100 credentialing organizations throughout the United States. There have been several subsequent releases that both extend and refine the vocabulary. The CTDL Schema Terms, Namespace Policy, Schema Serializations, and Guide are all available for public use under an open license from the *Credential Engine's* technical planning website (see listing of relevant project URLs in the Resources section of this text).

[2] http://credreg.net/ctdl/terms
[3] http://credreg.net/ctdlasn/terms
[4] http://standards.asn.desire2learn.com/
[5] http://shex.io/shex-primer/
[6] https://www.w3.org/TR/shacl/

CTDL development needed to satisfy two requirements: (1) sufficiently rich description of credentials and their context to support meaningful comparisons based on key factors such as cost, preparation time to award; locations, levels of quality assurance and specified competencies; and (2) the means for creating lightweight descriptions sufficient to support discovery. The first requirement is loosely illustrated in Figure 1. There are myriad ways in which a credential may relate to assessments and learning opportunities depending on certain conditions. For example, cost of a learning opportunity may depend on the type of person seeking the credential such as an in or out of state resident, veteran or military dependent. Costs may also vary across the regions in which it is offered. The Condition Profile makes it possible to capture these varying conditions. The second requirement is also reflected in Figure 1 by removing the Condition Profile entity and allowing instances of the credential class to make direct assertions about assessments, learning opportunities and competencies.

The presentation includes a review of the CTDL resources and development processes and will demonstrate how the CTDL schema is used for publishing RDF metadata to the Credential Engine Registry (CER), how an application profile of the CTDL is used by the CER to validate the quality of incoming metadata, and how the open application marketplace can evolve by demonstrating the Workit Search App prototype that consumes metadata from the CER. The presentation will include discussion of the planned development of CTDL-Lite as an addition to schema.org.

Resources

Credential Engine Informational Site: http://credentialengine.org

Credential Engine Technical Website: http://credreg.net/

Technical Advisory Committee http://credreg.net/tac

CTDL Schema: http://credreg.net/ctdl/terms

CTDL Guide: http://credreg.net/ctdl/handbook

CTDL Mapping Reference: http://credreg.net/ctdl/mapping

CTDL ASN Schema: http://credreg.net/ctdlasn/terms

CTDL ASN Terms (competency frameworks): http://credreg.net/ctdl/frameworkschemahowto

Credential Registry Guide: http://credreg.net/registry/handbook

Registry Assistant API Guide: http://credreg.net/registry/assistant

Credential Registry Search Query Builder Tool http://credreg.net/registry/search

Use Cases http://credreg.net/tac/usecases

Workit Search App Prototype (December launch): http://credentialfinder.org

VitroLib: From an Ontology and Instance Editor to a Linked Data Cataloging Editor

Presentation

Huda Khan
Cornell University, USA
hjk54@cornell.edu

Lynette Rayle
Cornell University, USA
elr37@cornell.edu

Rebecca Younes
Cornell University, USA
rebecca.younes@cornell.edu

Keywords: linked data; semantic application; cataloging editor

Abstract

The Mellon Foundation-funded Linked Data For Libraries Labs (LD4L Labs) and Linked Data For Libraries Production (LD4P) projects are exploring how to support library systems transition to the use of linked open data. As part of this work, we are developing a linked data cataloging editor called VitroLib. VitroLib extends Vitro, the open source ontology and instance editor that provides the ontology-agnostic semantic application underpinning VIVO, the researcher profiling system. VitroLib generates content display and content editing interfaces based on BIBFRAME, Bibliotek-o which extends BIBFRAME, and related ontologies. In this presentation, we will provide an overview of the design and implementation of VitroLib, results of usability testing exploring how catalogers can use VitroLib to catalog bibliographic metadata, and how VitroLib development has used application profiles.

We are utilizing a user-centered design approach to examine the needs of catalogers who are the target end-users for this application. VitroLib development includes the following main areas: (a) prototyping and evaluating the user interface for use by catalogers, and (b) ensuring the Bibliotek-o system of ontologies is expressed in the application correctly and according to the expectations encoded within the application profiles which are being developed concurrently. To understand cataloging workflows and how catalogers currently perform their cataloging tasks, we have had discussions and conducted usability testing with catalogers. The preliminary set of results identified the importance of searching for existing information even in the context of original cataloging and highlighted the incorporation of external vocabularies as a promising area of exploring the benefits of linked data. We intend on conducting further rounds of usability testing with future versions of the application.

Ensuring VitroLib adequately and correctly translates and expresses the Bibliotek-o system of ontologies in the interface requires more than simply incorporating ontology files. The challenges in this area result from both the Vitro software's current implementation and from how an ontology may not explicitly codify all the expectations for the properties defined. Vitro, as it currently comes out of the box, displays properties automatically only if they are OWL properties with specified domains or if the application is configured to display a property within certain contexts. Application profiles can provide useful context around the intended user interactions with the content as modeled by the ontology. Application profiles can help specify which classes are expected, even if not explicitly stated within the ontology, to be used for objects or subjects of a particular property. Additionally, profiles can specify which controlled vocabularies need to be used for a particular property. Part of VitroLib development is thus exploring the questions that the application profile creators can review and then implementing the application profile within the software confines of VitroLib.

Topic Maps for Digital Scholarly Monographs
Presentation

Alexandra Provo
New York University, USA
alexandra.provo@nyu.edu

Michel Biezunski
Infoloom, Inc, USA
mb@infoloom.com

Keywords: topic maps; resource discovery; digital scholarly monographs; EPUBs; linked open data; JSON-LD; subject metadata

Abstract

This presentation will outline work on a new approach to digital scholarly monograph subject metadata currently being undertaken by New York University's Digital Library Technology Services department as part of the Mellon-funded grant project, Enhanced Networked Monographs (ENM).

Expanding on the current NYU Press Open Access Books initiative (NYU Press, 2017), the ENM project will provide enhanced web-based access to selected books from NYU Press, the University of Minnesota Press, and the University of Michigan Press. In addition to annotation and full-text search, the ENM website will provide users with innovative paths of discovery and navigation via a "topic map" of names and concepts derived from back-of-book index entries.

The ENM project does not make use of the Topic Maps standard (ISO, 2006), but rather its basic underlying data model. ENM topic records are created and managed in the Topic Curation Toolkit (TCT). Developed by Infoloom Inc (Infoloom, 2017), the TCT is comprised of a database of topic records and a web-based editor interface. Through ingest scripts and subsequent human intervention, the TCT combines automatic and manual processing of EPUB indexes. NYU and Infoloom's experimentation with topic records as an alternate type of subject metadata represents a novel attempt to make machine use of an existing information structure hitherto geared principally toward human consumption.

This presentation will discuss the opportunities and issues that arise when transforming metadata created for the small-scale context of an individual book to a larger scale that cuts across texts. An overview of the workflow for creating topic map records, including a discussion of the role of human curation, will be provided. Topic map publishing will also be described, specifically the mapping of topic map records to JSON-LD and the design of ENM website features that make use of topics to provide users with enhanced navigation.

References

Infoloom, Inc. (2017) About Infoloom. Retrieved June 12, 2017, from https://www.infoloom.com/who-we-are/

ISO. (2006). ISO/IEC 13250-2:2006: Information technology -- Topic Maps -- Part 2: Data model. Retrieved June 12, 2017 from https://www.iso.org/standard/40017.html

NYU Press. (2017) Open Access Books: About this Project. Retrieved June 12, 2017, from http://openaccessbooks.nyupress.org/about/

Posters

Integrated Learning of Metadata Quality Evaluation and Metadata Application Profile Development in a Graduate Metadata Course

Poster

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

Abstract

This report describes an experiment in the design of an advanced graduate metadata course to facilitate more efficient link between content-based learning and skill-based learning. The experiment included integrating the process of designing a local metadata application profile with learning evaluation of metadata quality, including leaning to assess the ability of a standard metadata scheme or an application profile to capture and adequately represent important and unique attributes of information objects in a special collection. The benefits of this approach are discussed.

Keywords: metadata education; metadata application profiles; metadata evaluation; metadata quality.

1. Introduction

The landscape of metadata work has changed dramatically in the recent two decades and continues to rapidly evolve. Ability and willingness to learn and flexibility are now among the most often required traits for metadata specialists. As a result of the shift to knowledge-based economy, one of the two integral components of knowledge – skills – needs to receive more emphasis in designing the educational programs than previously when education was more content-focused. Professional associations such as Association for Library Collections and Technical Services publish information on the skillsets for metadata professionals. The skills that the employers are looking for in metadata specialists have been examined by the studies analyzing job ads and other related materials (e.g., Hall-Ellis, 2006; Han & Hswe, 2010; Park & Lu, 2009). Surveys of metadata practitioners and metadata educators (e.g., Hider, 2006; Hsieh-Yee, 2004; Park & Tosaka, 2010) identify metadata quality evaluation skills as one of the priorities in metadata education. Several in-depth case studies (e.g., Glaviano, 2000, Hsieh-Yee, 2000; Or-Bach, 2005) contribute to understanding of how the metadata skills are developed through assignments and other course activities. However, none of them focused on the skills of growing importance: metadata quality evaluation and metadata application profiles development. We attempt to address this gap in the project briefly presented below.

2. Course Design for Learning Metadata Quality and Application Profiles

The University of North Texas (UNT) graduate students are offered a selection of six metadata courses. Four graduate courses focus on various aspects of Machine-Readable Cataloging (MARC) metadata and/or classification systems used in libraries. The remaining two graduate courses represent a sequence of an introductory and advanced metadata courses. Students take the advanced metadata course after completing the introductory course in which they learn about the structure of metadata schemes, metadata elements, semantics, and syntax, familiarize through readings and practice with the use of HTML and XML in metadata records, develop theoretical and practical understanding of Dublin Core, Metadata Object Description Schema (MODS), and Visual Resources Association's VRA Core 4.0 metadata and one of the existing metadata

application profiles:(Dublin Core Collections Application Profile (DCCAP). The more complex topics such as principles guiding creation of metadata application profiles (MAPs) and the process of building MAPs, along with other important advanced topics such as metadata quality, metadata interoperability, and expression of metadata as Linked Data, are covered in the advanced metadata course.

The advanced metadata course design experiment includes close integration of course topics through the sequence of assignments, in which the work students completed as part of one assignment informs the work completed in the next assignment. The content-based and skills-based learning on the topic of MAPs are separated in time. The content knowledge is delivered early in the semester, when students learn about MAPs through instructor's lectures and required readings, and discuss their understanding of the principles guiding MAPs development and characteristics of existing MAPs in the course discussion forums. The MAPs learning module is followed by the metadata quality learning module, and the learning module on metadata in digital content management. In the major assignment which culminates the semester, the MAPs skill building is integrated with content knowledge and skills obtained in all learning modules.

The experiment mainly focuses on two major assignments: Metadata Evaluation and Documentation, and Metadata Application Profile. In the Metadata Evaluation and Documentation assignment, each student collects their own two small random samples of metadata records (created according to a local version of qualified Dublin Core) from two digital collections available through the Portal to Texas History: a baseline collection and a target collection. Students analyze metadata quality in these records in relation to the major criteria of completeness, accuracy, and consistency (as defined by e.g., Bruce & Hillmann, 2004; Moen, Stuart, & McClure, 1998), both within each sample and comparatively across the two samples; and write a summary of comparative metadata evaluation results. After completion of metadata evaluation tasks, students draft metadata creation guidelines for the target collection. This task is informed by student metadata evaluation findings, as well as their understanding of the specific attributes of information objects in the collection and the ability of the given metadata scheme to accommodate representing these attributes.

In creating metadata documentation, students use as the starting point the existing guidelines for the baseline collection. The process involves categorizing information in the existing guidelines document into three categories: applicable to representing objects in the target collection, conditionally applicable, and those completely inapplicable. The criteria used in selecting a baseline collection included students' familiarity with the collection, availability of detailed collection-specific metadata creation guidelines, and collection homogeneity. All students in advanced metadata course had previous exposure to this homogenous collection which consists of a single type of information objects (patents) in the capacity of metadata creators through one of the exercises in the introductory metadata course. In the process of creating metadata records in the introductory course, students developed understanding of patents and gained familiarity with metadata guidelines for collection.

The criteria for selecting a target collection include homogeneity of collection, the absence of collection-specific metadata creation guidelines, and the mostly visual nature and short content length of items which expedites the process of evaluating information objects and representing them with metadata. In the initial course offerings, a collection of postcards had been used but later the institution contributing collection to the aggregation expanded the collection scope by including other types of information objects. As the postcard collection lost its homogeneity, determining typical collection-specific item attributes and meaningfully comparing this collection with a homogenous patent collection became impossible. For that reason, in the latest iteration of the course design a collection of architectural drawings was used as a target collection. Architectural drawings were deemed the optimal point of comparison to patents for the course exercise purposes due to similarities between the two types of information objects.

In the previous offerings of the course, students completed the readings and discussions of the MAPs and the practical exercise closer to the beginning of the semester: prior to learning about metadata quality, evaluating metadata quality and developing metadata creation guidelines. Each student was assigned their own small collection to design a metadata application profile for. This approach, however, was found to lack continuity: while the students developed understanding of and interest in MAPs and other advanced metadata topics, their ability to clearly see the connections between these topics – especially the connections between the design of a MAP on one side and developing the guidelines for metadata creators and evaluating the quality of resulting metadata on another side – was not adequately supported by the practice. Therefore, the decision was made to more closely integrate the course topics through the sequence of assignments, in which the work completed as part of one assignment would inform the work completed in the next assignment and the issues encountered in the later assignment would give students a chance to reflect to the topic of the earlier assignment. In the most recent iteration of the course, students complete the Metadata Application Profile assignment at the end of the semester, after having developed the content knowledge on MAPs and other course topics and having completed and received instructors' feedback on other skill-building practical assignments, including the Metadata Evaluation and Documentation exercise. In designing their own MAP for architectural drawings, students build on their previous work on evaluating architectural drawings, as well as on assessing the suitability of a specific MAP – used in the Portal to Texas History for describing materials in all collections – for represent architectural drawings.

The latest version of Metadata Application Profile assignment consists of three parts. In the first part, students estimate the target audience for architectural drawings and compile a list of the attributes that will likely be of importance to the target audience. Next, students introduce the metadata elements to represent these attributes, and provide specifications for each element. The minimum requirement for the MAP is to consist of at least 17 metadata elements, including each of these three categories:

- applicable for describing architectural drawings existing metadata elements adopted from two (2) or more standard metadata schemes, including but not limited to Dublin Core,
- existing element(s) adapted – with modifications for representing the architectural drawings – from standard metadata schemes, and
- new local metadata element(s) defined by students.

Students make decisions on definitions, vocabulary control and cardinality of each metadata element, as well as on mapping to standard elements, on the order of elements in the record. They are also asked to express the element names with namespace the way they would be expressed in a DSpace-powered digital repository, using the solutions to overcome the problems with DSpace accommodation of hierarchical metadata schemes (based on what students learned in another learning module in this course).

In Part 2, students express the MAP design ideas resulting from completing Part 1 as a data model in the RDF/XML. For this task, students are instructed to use as a template a copy of the 2012-06-14 release of DCMI Metadata Terms data model (http://dublincore.org/2012/06/14/dcterms.rdf). In the final section of the assignment, students test the resulting MAP. For this purpose, students follow the specifications of their own MAPs to create a metadata record describing one familiar architectural drawing from the sample analyzed in Metadata Evaluation and Documentation exercise. This allows to see the connection between theorizing and implementing a specific MAP, to discover practical problems with implementing the MAP designed in Part 1 and Part 2 and to fix them as needed based on the test results.

To support the learning, a substantial amount of time in the weekly class meetings is devoted to discussion of student ideas for and challenges in the process of the architectural drawings MAP development and the ways in which their design is informed by their findings on existing

metadata quality in the collection of architectural drawings. Course discussion forums are also extensively used for discussing these issues and establishing connections between MAP design and implementation and developing and using guidelines for metadata creators, and importance of metadata quality evaluation.

3. Conclusions

The redesign of advanced metadata course described in this report was based on the assumption that revisions would improve the overall quality of learning – and skill-building in particular – as well as student satisfaction. The data collected as part of this experiment allows to make the conclusion that this assumption was correct. The average quality of student work in both Metadata Evaluation and Documentation exercise and Metadata Application Profile exercise (as expressed in assignment submissions assessment) has improved after implementing the course design change; this improvement was the most noticeable for metadata quality learning: from 89.6 to 94.88 out of 100. The average quality of student learning on these and other advanced metadata topics, as evaluated by the teaching team and expressed in semester grades, has also improved. Student evaluation of the course also shows the benefits to the quality of learning. Student perception of two indicators – (1) usefulness of written (skill-building) assignments in understanding of the course content, and (2) overall course quality – has substantially improved (by 9.52%).

When detailed description and results of this project are published, we expect this will make a contribution to understanding of efficient approaches to developing crucial skills in the process of providing graduate education to metadata professionals. Hopefully, it would encourage other course developers and instructors of metadata courses to share their best practices. The platform for sharing these ideas and learning objects would be very beneficial in improving the quality of metadata education.

References

Bruce, T.R., & Hillmann, D.I. (2004). The continuum of metadata quality: defining, expressing, exploiting. In Hillman, D. and Westbrook, L. (Eds.), Metadata in Practice. Chicago: American Library Association, pp. 238-256.

Glaviano, C. (2000). Teaching an information organization course with Nordic DC metadata creator. *OCLC Systems & Services: International Digital Library Perspectives, 16* (1).

Hall-Ellis, S. D. (2006). Cataloging electronic resources and metadata: Employers' expectations as reflected in American Libraries and AutoCAT, 2000-2005. *Journal of Education for Library and Information Sciences, 47* (1), 38-51.

Han, M.J., Hswe, P. (2010). The evolving role of the metadata librarian. *Library Resources and Technical Services, 54* (3).

Hider, P. (2008). A survey of continuing professional development activities and attitudes amongst catalogers. *Cataloging & Classification Quarterly, 42*(2), 35-58.

Hsieh-Yee, I. (2000). Organizing Internet resources: teaching cataloging standards and beyond. *OCLC Systems & Services: International Digital Library Perspectives, 16* (3).

Hsieh-Yee, I. (2004). Cataloging and metadata education in North American LIS programs. *Library Resources and Technical Services, 48* (1), 59-68.

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

Or-Bach, R. (2005). Educational benefits of metadata creation by students. *ACM SIGCSE Bulletin, 37* (4), 93-97.

Park, J.R., & Lu, C. (2009). Metadata professionals: roles and competencies as reflected in job announcements, 2003–2006. *Cataloging & Classification Quarterly, 47*(2), 145-160.

Park, J.R., & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly, 48*(8), 696-715.

The Development of Application Profile for OAK Institutional Repository

Poster

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

Jee-Hyun Rho
Pusan National University, South Korea
jhrho@pusan.ac.kr

Eun-Ju Lee
Dongeui University, South Korea
ejulee@deu.ac.kr

Keywords: institutional repository; metadata; OAK; application profile; Dublin Core; DSpace

Abstract

OAK (Open Access Korea) hosted by National Library of Korea is the national portal for integrated search of IRs, participating universities, public institutions, researches, and businesses. OAK had collected metadata from member IR systems, and has accumulated them according to the OAK metadata scheme by mapping IRs metadata into OAK metadata scheme. OAK has developed OAK metadata scheme based on DSpace to build OAK portal in 2009, but the initial OAK metadata schemes could not accommodate all the elements that participant IRs wanted. Elements of member IR metadata were redundant, disorganized and inappropriate. As a result, it brought into missing metadata in harvesting and mapping the metadata from member IR systems, and into searching without the elaborate elements. For solving these problems, this study is to suggest the OAK application profile through developing new OAK metadata scheme by accommodating member IRs metadata and comparing elements between OAK metadata and major IRs such as DSpace, Eprints, BEPress, ETD-db, and dCollection sophisticatedly. As the research methods, it is used to analyze the 17 representative IRs' metadata schemes and compare the major IRs' metadata schemes for conforming standard metadata scheme.

Analyzing the Metadata Elements of 17 IRs

The metadata schemes of 17 among 33 OAK IRs were analyzed in aspect of 15 main elements of Dublin Core.

In title, various titles were used as sub-elements such as translated title, original title, subtitle, part title, and part number. Subject element was qualified into sub-elements for differentiating local classification schemes and various subject headings vocabularies. In description, claim (request for a patent), version (peer-reviewed), and provenance (owner of resource) were used as the element refinement. In contributor, various contributors were used with sub-element such as editor, illustrator, examiner, googleAuthor, college, department as well as author, advisor. Date was divided into element refinements such as created, available, issued, submitted, accessioned, updated, valid, and modified. In type, various type encoding schemes were used including MARC type, DSpace type, and so on. So, the mapping table between resource type schemes should be made to retrieve the resources by resource type. In identifier, it was not affordable to accept all the identifiers such as uri, govdoc, isbn, ismn, issn, sici, patentno, pmid (pubmed identifier), scopusid (scopus identifier). In citation, citation description methods were various in IRs -- one IR subdivided citation into several element qualifiers such as title, volume, date, genre, identifier, issueno, and page, other IR described the citation information under the relation element, and another IR moved the citation to the main element not as the sub-element of identifier. The citation was the element in need to uniformity.

There are some difficulties in retrieving the resources in national portal level because each IR has used its own elements and sub-elements. Therefore, new OAK metadata scheme should be developed to consolidate all the IRs metadata elements.

Comparing the major IRs' metadata

The representative IRs' metadata schemes in the world were selected for comparison of metadata elements such as DSpace, Eprints, BEPress, ETD-db, and dCollection (IR for college and university libraries in Korea, host by KERIS). <TABLE 1> shows the comparison of only title and description sections of 5 IR metadata schemes.

TABLE 1: Comparison table of title and description sections of 5 IR metadata schemes (in part).

DSpace		EPrints	BEPress	ETD-db		dCollection	
element	qualifier			element	qualifier	element	qualifier
title		title	title, article title	title		title	
	alternative	alternative title			alternative		alternative
							subTitle
							translated
description		description		description		description	
	tableofcontents						tableofcontents
	abstract		abstract		abstract		abstract
	provenance	provenance					provenance
	sponsorship	funder					sponsorship
		grant number					
		parent project					
	statementofresponsibility						
	uri						descriptionURI
			embargo period				
			versions		release		
			peer reviewed				indexed
		publication status	publication status				
		data collection method					
		contact					
		administrative note			note		
		additional information					
				degree	name		
		divisions*	disciplines		discipline		
					grantor		
					level		localRemark
			comments				

This table was based on Chung, Yeon-Kyoung, Na-Nee Lee and Mihaw Lee (2007), KERIS (2005), BEPress Home Page, DSpace Home Page, Ensom, Tom and Alexis Wolton (2012), Jones, Richard (2004), Digital Commons@Otterbein.

EPrints includes special elements and sub-elements to reflect the project related information such as funder, grant number, and parent project. BEPress has specialty in describing journal article, for example, version (preprint, postprint, published), publication status, peer reviewed, and embargo period. ETD-db develops thesis related element such as degree, name, level and discipline. dCollection includes alternative, subtitle, and translated in title, and indexed as the sub-elements of description. Also, dCollection has the various sub-elements in identifier.citation to build the citation information - citationTitle, citationIdentifier, citationGenre, citationOIssueNo, citationVolume, citationNumber, citationTissueNo, citationPages, citationStartPage, citationEndPage, citationConferenceNumber, citationConferencePlace, citationConferenceDate,

citationEdition, citationAuthor, place, duration, publisher, uriType, uriEntity, ISBN, ISSN, SICI, KERISIdentifier.

It is to find that OAK metadata scheme should have been developed with the various elements and sub-elements through analyzing the world wide IRs metadata schemes. These elements and sub-element should be applied to OAK metadata schemes.

Development of New OAK Application Profiles

The objective of this study was to suggest new OAK application profiles by analyzing the elements and sub-element of member IRs metadata and comparing major international IRs. The previous OAK metadata scheme was changed -- new elements and sub-elements were added, and elements with similar meaning were integrated into one element or sub-element. The new OAK application profile's features are as follows:

- The different elements which have same meaning are unified to one element or sub-element, for example, in description, abstract, summary as sub-elements → summary.
- Subject, degree, eprintVersion, contributor, nameIdentifier, and identifier are changed to subjectType, degreeType, eprintType, contributorType, nameIdentifier, and identifierType to select type from controlled vocabularies to accommodate various elements of member IRs and to differentiate the value. For example, it is possible to differentiate various contributor such as author, advisor, editor, translator, illustrator, examiner, department, reviewer by inputting type from controlled vocabularies of contributorType.
- New elements and sub-elements are added such as subject.keyword (keyword written by author), and description.degree (degree type).
- According to Dublin Core Metadata Initiative Citation Working Group (2005), bibliographicCitation is to capture the bibliographic citation information for a resource, but, bibliographicCitation is not enough to construct and describe the citation information in uniform. Therefore, in OAK, the citation element was changed as the main and administrative element and was subdivided into citation.title, citation.volume, citation.number, citation.date, citation.startPage, citation.endPage, citation.conferenceName, citation.conferenceNumber, citation.conferenceDate, citation.author, citation.author, citation.edition, citation.place, citation.publisher to embrace all kinds of resource's citation information. These citation related elements are used only as administrative element to get the data from input interface, and identifier.bibliographicCitation was used as element in displaying the citation information by collecting data in citation such as Table 2.

TABLE 2: Citation description example

Property	Value String	notes
OAK:title	A critique of the FRBR user task and their modifications	
OAK:contributor (value=author)	Hider, Philip	
OAK:publisher	Taylor & Francis	
OAK:identifier.bibliographic Citation	Cataloging and Classification Quarterly, 55(2), 55-74.	This data is displayed for the citation information.
OAK:citation.title	Cataloging and Classification Quarterly	This data is used only for getting the data from user.
OAK:citation.volume	55	
OAK:citation.number	2	
OAK:citation.date	2016	
OAK:citation.startPage	55	
OAK:citation.endPage	74	

Future work

This study is to make OAK AP to accommodate all member IR metadata according to international standard such as DC, DSpace et al. This OAK AP would contribute to uniformity and consistency of portal metadata. After system development using new OAK metadata scheme, user survey and evaluation must be completed as the future works.

Acknowledgements

This paper was funded by National Library of Korea.

References

BEPress Home Page. Retrieved, September 20, 2016, from http://www.bepress.com.

Chung, Yeon-Kyoung, Na-Nee Lee and Mihaw Lee. (2007). A study of an extension of metadata for institutional repository. Journal of the Korean Society for Library and Information Science, 41(1), 323-344.

Digital Commons@Otterbein. Retrieved, September 20, 2016, from http://digitalcommons.otterbein.edu/metadata_elements_master_list.pdf.

DSpace Home Page. Retrieved, September 20, 2016, from http://www.dspace.org.

Dublin Core Metadata Initiative Citation Working Group. (2005). Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata. Retrieved, September 20, 2016, from http://dublincore.org/documents/dc-citation-guidelines.

Ensom, Tom and Alexis Wolton. (2012). RDE Metadata Profile for EPrints. Retrieved, September 20, 2016, from http://www.data-archive.ac.uk/media/375386/rde_eprints_metadataprofile.pdf.

Jones, Richard. (2004). DSpace VS. ETD-db: Choosing software to manage electronic theses and dissertations. Ariadne, 38. Retrieved, September 20, 2016, from http://www.ariadne.ac.uk/issue38/jones.

KERIS. (2005). dCollection Manual for System Managers. Seoul: KERIS.

Facilitating Information Sharing and Collaboration through Taxonomy at the Federal Reserve Board

Poster

Jennifer Gilbert
Board of Governors of the Federal Reserve System, United States
jennifer.g.gilbert@frb.gov

Alison Raab Labonte
Board of Governors of the Federal Reserve System, United States
alison.r.labonte@frb.gov

Franz Osorio
Board of Governors of the Federal Reserve System, United States
franz.osorio@frb.gov

Keywords: data inventory; expert directory; facets; metadata; taxonomy

Abstract

The Research Library at the Board of Governors of the Federal Reserve System developed the Board Subject Taxonomy (BST) by organizing and standardizing key concepts into a vocabulary of subject terms that describe staff economists' research and policy work. The goal was not just to have a taxonomy; rather, we sought a way to better facilitate sharing, collaboration, and discovery of information across systems. To that end, the Library staff has developed several tools to allow the taxonomy to forge relationships and connections across disparate sources. The BST acts as a critical semantic link to bring together data, researchers, and publications that were previously isolated from each other. The BST is currently deployed in a data inventory (*DataFinder*), research publication repository (*OneBoard Research*), an expert directory (*Board Expert Finder*), and a researcher index (*Economist Similarity Index*). The Board Subject Taxonomy is significant in that it brings together economists' research and interests using the Federal Reserve vernacular, to help transcend the silos of information in our agency. The BST is central to metadata quality as it helps keep the tools we developed in sync with each other and produces interoperability.

The BST was developed in accordance with controlled vocabulary standards and is influenced by a range of taxonomies, from the Journal of Economic Literature (JEL) Classification System and the Thesaurus for Economics (STW) created by the Leibniz Information Centre for Economics (ZBW) to internal lists of keywords. The project was initiated in late 2012 by a senior management directive to provide a uniform list of terms to describe Board research across the various divisions. Initial work involved focus groups consisting of researchers and policy analysts, surveys, and evaluation of existing vocabularies in use at the Board. Next, prospective vocabularies were developed by merging JEL with internal, home-grown vocabularies and mining keywords and tags associated with our economists' research. A beta version of BST was released in April 2013, and terms began to be applied to working papers by 2014. In late 2014, the team shifted the management of the terms from a spreadsheet to a thesaurus management tool that could help keep relationships among the terms up to date and impose formatting based on ANSI/NISO Z39.19-2005. The change in tools moved the taxonomy from a flat list to a hierarchical structure with an emphasis on BT/NT and RT relationships. Beginning in late 2016 and continuing into the near future, focus is on the transformation of the thesaurus from a single hierarchy to a faceted taxonomy, to aid in search and discovery, and to meet the need for esoteric vocabularies used at the Board. Facetization began with the identification of branches within the taxonomy that could answer user questions on their own. These were concepts such as types of financial institutions and instruments. The creation and maintenance of the BST has served as the Research Library's entry to the Board's initial linked data efforts. Recent work has focused on supporting the transformation of the taxonomy into RDF for use in a semantic search tool.

Discovering relevant terminology from each operational unit to build a robust taxonomy was, and in some cases still is, hindered by the lack of a shared vocabulary. Our librarians worked with

users in various sections of the Board to gain an understanding of their vocabularies/classification systems and information organization practices. This partnership was essential to establish the BST, and the collaboration continues. Key technological challenges that we have addressed: 1. the means by which the taxonomy is managed; and 2. the means by which it is applied. Currently, staff manually assign and maintain terms. In the future we aim for automation in term assignment and management.

The real measure of our success has been the myriad ways we have used the BST to improve search and discovery at the Board. These include:

- OneBoard Research, a digital repository for finding Board research tagged with BST terms;
- DataFinder, an online inventory of licensed and acquired data assets tagged with BST terms;
- Board Expert Finder, a tool for locating subject experts, to help build communities of practice. BST terms assigned to publications in OneBoard are used to generate a directory of experts;
- Economist Similarity Index, an index for assessing research interests, thus helping Board researchers find and make new connections with individuals having similar interests. This last endeavor is one of the more exciting, because it emerged from end user conversations and is entirely responsive to user needs.

Our key accomplishment has been to exploit the power of robust metadata: using it to repurpose information in diverse systems (library catalogs/inventories, research repositories) and collect that information for use in home-grown, dynamic search tools that meet user needs. For the development of the Economist Similarity Index, to give an example, we worked with a user who wanted to find economists who wrote on similar subjects. Because we use the same taxonomy to describe all publications, we leveraged that metadata to answer the question and built an Index based on an algorithm had similarity ratios for each economist

In the future, our efforts will be based on user demands and have strong connections to business needs at the Federal Reserve Board. These efforts are to include:

- Expansion of vocabulary and facets to include terms relevant to collected data (that is, data collected by the Board in its supervisory and regulatory role);
- Development of user-specific vocabularies for Board Information Technology specialists;
- Establishment of a Diversity vocabulary, in line with the Strategic Mission of the Federal Reserve Board, to illustrate the depth of research done on economic inclusion;
- Expansion of terms to describe the technical and mathematical aspects of economics research more granularly;
- Repurposing of our metadata, including BST in an enterprise-level data inventory, powered by a semantic search tool. This coincides with our move towards revising the BST using RDF/SKOS, which offers greater interoperability;
- Marketing and branding of the taxonomy to raise awareness of it as a support tool that aids in search and discovery;
- Opening up OneBoard Research, the Board's research repository, on the public website;
- Providing the BST as linked open data so other institutions/individuals can make use of it.

References

American Economic Association. (2017). JEL Classification Codes Guide. Retrieved August 30, 2017, from https://www.aeaweb.org/jel/guide/jel.php.

German National Library of Economics (ZBW). (2017). STW Thesaurus for Economics. Retrieved August 30, 2017, from http://zbw.eu/stw/version/latest/about.

NISO. (2014). ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Retrieved August 30, 2017, from http://www.niso.org/apps/group_public/project/details.php?project_id=46.

ORCID: Using API Calls to Assess Metadata Completeness

Poster

Naomi Eichenlaub
Ryerson University, Canada
neichenl@ryerson.ca

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

Keywords: ORCID; persistent identifiers; researcher identifiers; interoperability; APIs; public data; name disambiguation; metadata assessment.

Abstract

The aim of this poster is to demonstrate the importance of adequate metadata in ORCID profiles to ensure name disambiguation. It is only through more complete metadata that ORCID will ensure success in terms of interoperability with institutional scholarly, publishing and funding bodies.

Introduction

Launched in 2012, ORCID (Open Researcher and Contributor ID) is a non-profit persistent digital identifier open registry offered to researchers across disciplines. Their mission is to provide an identifier in the form of a unique alphanumeric code to provide persistent identity for researchers. There are three easy steps to receive a persistent ID and distinguish oneself from other researchers: register, add your info, and use your ORCID ID. However, we noticed that a very large number of ORCID IDs are empty, i.e., a name is registered and a profile is created but they lack critical elements required to perform the function of name disambiguation such as country and affiliated institution. Other metadata, for instance activities summary, funding, peer-reviews, and works can be considered important but not critical to the primary function of ORCID, which is name disambiguation. ORCID itself has noted this problem in a blog post in early 2017 stating that "Many records have limited (public or trusted party) information beyond the author's name and iD" (ORCID, 2017a). Duplicate records (created when an author forgets they have already created an ORCID) are a much smaller issue and are being addressed by ORCID with new initiatives such as self-management of duplicate records (ORCID, 2017a). Essentially, the richer the metadata the more useful something becomes so it behooves the research community as a whole to work together to make ORCID records as rich and complete as possible. Therefore, in an effort to get a better sense of the overall completeness of the current state of ORCID records, we decided to investigate and query the records and metadata fields using the public ORCID API.

Methodology and Results

We ran queries against the public ORCID API to get a better sense of how many ORCID records have only minimal information. In order to query the public file, we registered for a public API client application, enabled the developer tools for the application, and accessed the authorized URL to retrieve an authentication code. Once we accessed the token URL, we retrieved an authentication token, which was used for all the API calls. To invoke RESTful API calls we used Postman software that enabled the following headers: Accept: application/json, Authorization_type: Bearer, and the Access_token.

To get a sense of the big picture in terms of records with only the minimal required information, i.e., name and email, we searched records created between 2012-2017 that did not

include affiliation / organization name, Ringgold ID, and any work titles. Results were gathered by year as well as an overall API call to search the entire ORCID database.

TABLE 1: Examples of public API calls.

Year	API Calls	Count
2012	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st to December 31st, 2012: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2012-01-01T00:00:00Z%20TO%202012-12-31T00:00:00Z%5D	25,351
2013	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st to December 31st, 2013: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2013-01-01T00:00:00Z%20TO%202013-12-31T00:00:00Z%5D	258,182
2014	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st to December 31st, 2014: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO*]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2014-01-01T00:00:00Z%20TO%202014-12-31T00:00:00Z%5D	370,074
2015	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st to December 31st, 2015: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2015-01-01T00:00:00Z%20TO%202015-12-31T00:00:00Z%5D	479,144
2016	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st to December 31st, 2016: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2016-01-01T00:00:00Z%20TO%202016-12-31T00:00:00Z%5D	709,046
2017 to 05/17/17	ORCID records with any given names, without any affiliation, Ringgold ID, work titles, and submitted between January 1st, 2017 to May 17, 2017: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]+AND+profile-submission-date:%5B2017-01-01T00:00:00Z%20TO%202017-12-31T00:00:00Z%5D	372,709
2012 to 05/17/17	All ORCID records with any given names, without any affiliation, Ringgold ID, and work titles: https://pub.orcid.org/v2.0/search/?q=given-names:[* TO *]+AND+-affiliation-org-name:[* TO *]+AND+-ringgold-org-id:[* TO *]+AND+-work-titles:[* TO *]	2,216,944

The API calls were made on May 17, 2017. The results as seen in the above table are relevant to that date and may have changed since then. Current ORCID statistics can be viewed at the following URL: https://orcid.org/statistics.

Moreover, based on the above results and the total number of ORCID records submitted between 2012 and 2017 (up to May 17, 2017) we calculated the percentage of minimal ORCID records in the respective year. More than half of the records submitted are minimal.

TABLE 2: Percentages of minimal ORCID records by year.

Year	Total ORCID Submissions	Minimal ORCID Records Count	Percentage
2012	44,118	25,351	57.46%
2013	424,927	258,182	60.75%
2014	608,999	370,074	60.76%
2015	769,979	479,144	62.22%
2016	1,049,820	709,046	67.53%
2017*	948,445	372,709	39.29%

* Up to May 17, 2017

Challenges

Missing data such as affiliations or work titles poses a major challenge on one hand to gather appropriate data and on the other to consider PID service adoption. From the preliminary results we noticed that the problem of orphan (empty) records is very common. Many records used random placeholder names such as John/Jane Doe to keep their true identity unknown, filler text such as lorem ipsum, fictitious funding, works, and institution names for employment and education.

At this stage in our research it is beyond our purpose to search how many names are similar and/or unable to be disambiguated due to lack of additional information even though technically not impossible.

In a blog post on the challenges of measuring PID (Persistent IDentifier) adoption by Robin Dasler (senior fellow in CERN's Scientific Information Service) she pointed out that ORCID acknowledges that when the service first launched, "it was fine to be concerned only with uptake, since the priority was to get the word out." However, with the growth and development that has occurred over the years, the focus needs to be on attaining innovation and interoperability (ORCID, 2017b).

In an effort to foster integration and engagement within the research community, ORCID launched the Collect & Connect program in 2016. With increased trust in connections "between researchers and their professional affiliations and activities" a greater number of ORCID identifiers can be collected and connected (Meadows, A., 2016), thereby maximizing metadata robustness and interoperability (ORCID, 2017c).

Conclusions and Future Work

ORCID IDs are very useful, specifically when most names are not unique. However, many records lack critical elements required to perform the main function of a personal identifier, namely the name disambiguation.

To this end, the authors propose that one of the priorities going forward should be to work together to ensure that a greater number of ORCID records have a higher number of completed metadata fields especially since incomplete metadata poses a challenge to name disambiguation. This issue has been acknowledged by ORCID as well: "We need to do more to ensure that ORCID identifiers are collected using appropriate, validated methods, and are published with research activities and affiliations" (Meadows, A., 2016).

In order to address a large number of ORCID records missing critical metadata fields the following solutions are proposed:

1. Advocacy and education - via their unique positions, librarians in academic settings can provide research support services such as individual and personalized researcher profile consultation services offered to targeted researchers at an institution in order to ensure ORCID records are more complete (Thompson, E. & French, S., 2017; Reed, K., McFarland, D., & Croft, R., 2016), and can lead "campus-wide efforts to promote the use of ORCID and similar resources" (Tran, C.Y. & Lyon, J.A., 2017).

2. Continued ORCID outreach: ORCID has identified a "key goal" for 2017 "to develop education and outreach resources for researchers explaining how and why to connect information in the ORCID records" (ORCID, 2017a).

3. Increased system interoperability: ORCID notes that they are "working with third party system vendors to improve information flow between systems" and that they will "continue to expand the types of works and activities that can be connected to ORCID records" (ORCID, 2017b).

It is hoped that this poster draws attention to the problems associated with a lack of identifying metadata in ORCID records and highlights the value of more complete metadata in disambiguating researcher names and identifiers.

References

Meadows, Alice. (2016). Everything you ever wanted to know about ORCID … but were afraid to ask. College & Research Library News, 77(1), 23-30.

ORCID. (2017a). ORCID open letter: One year on. Retrieved, May 15, 2017, from https://figshare.com/articles/ORCID_Open_Letter_-_One_Year_On_Report/4828312.

ORCID. (2017b). Challenges of measuring PID adoption. Retrieved, April 17, 2017, from http://orcid.org/blog/2017/03/24/challenges-measuring-pid-adoption.

ORCID. (2017c). Collect & Connect. Retrieved, March 15, 2017, from https://orcid.org/content/collect-connect.

ORCID. (2017d). ORCID API. Retrieved, March 15, 2017, from https://orcid.org/organizations/integrators/API.

Reed, Kathleen, Dana McFarland, and Rosie Croft. (2016). Laying the groundwork for a new library service: Scholar-practitioner & graduate student attitudes toward altmetrics and curation of online profiles. Evidence Based Library and Information Practice, 11(2), 87-96. doi:10.18438/B8J047.

Ringgold. (2017). Case Studies ORCID. Retrieved, March 15, 2017, from http://www.ringgold.com/case-studies-orcid.

Thompson, Elleb and Sally French. (2016). Pimp my Profile and the researcher profile health check: Practical, individualised researcher support initiatives co-created by library and faculty. In ALIA National Conference 2016, 29 August - 2 September 2016, Adelaide, SA. Retrieved, March 15, 2017, from http://eprints.qut.edu.au/98649/.

THOR. (2017). Researcher metrics: ORCID IDs. Retrieved, March 15, 2017, from http://dashboard.project-thor.eu/dashboard/researcher/.

Tran, ClaraY. and Jennifer A. Lyon. (2017). Faculty use of author identifiers and researcher networking tools. College & Research Libraries, 78(2), 171-182. doi:10.5860/crl.78.2.171.

Estimating Domain Models from Metadata Instances to Improve Usability of LOD Datasets

Poster

Ryota Kinjo
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
S1721665@s.tsukuba.ac.jp

Mitsuharu Nagamori
Faculty of Library, Information and Media Science, University of Tsukuba, Japan
nagamori@slis.tsukuba.ac.jp

Shigeo Sugimoto
Faculty of Library, Information and Media Science, University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp

Keywords: metadata; metadata schema; domain model; schema extraction; application profile;

Introduction

Linked Open Data(LOD), which is one of the efforts to help realize semantic web, has gradually become popular. Many Linked Open Data datasets, however are not well utilized. There are multiple reasons for this, such as limited recognition of LOD, limited usability of LOD datasets and so on. In attempting to solve these issues, we focused on a metadata schema that describes the structure about metadata instances in each LOD dataset. As information about metadata schema are not typically released, it is difficult to use LOD datasets. Therefore, in this research we extract the domain model (Dublin Core) as directed graph, which is one piece of information about a metadata schema, from metadata instances. For example, FIG1 illustrates a domain model extracted from a metadata instance.

Domain models are suitable for understanding the rough structure of a metadata instances in an early stage. We developed an estimation method to generalize a process of understanding metadata schema when people who are not familiar to the datasets handle them. We then apply the estimation method to existing datasets.

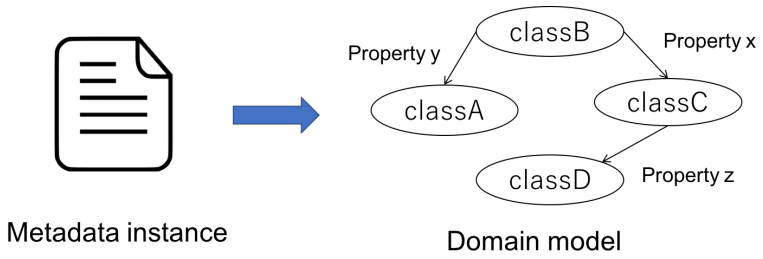


FIG. 1. Estimate domain model from metadata instances.

Method for Estimating Domain Models

People generally try to grasp the rough structure of datasets by executing SPARQL queries and seeing text formatted metadata. When people understand things and relation among the things in metadata, people can better understand metadata schema. Therefore, the purpose of a domain model that we estimate is to help people understand main class and properties which belong to those main classes. The method for estimating domain model is divided into 3 steps.

First is the execution of a series of SPARQL queries, as shown in FIG.2, to get statistics that are needed to estimate the domain model. Second is to determine which information is put into the domain model. This is the most important step in estimating the domain model. We implemented this through the comparing of each number of class as a subject with number of class as an object.

"Number of class as a subject" means number of times that a class behaves as a subject as shown in FIG.3. For example, if a resource which belongs to class A appears three times as a subject in metadata instances, the number of class A as a subject is three.

```
SELECT DISTINCT ?type
WHERE{
  ?s rdf:type ?type .
}
```

(a) A query to get all classes

```
SELECT DISTINCT ?v
WHERE{
  ?s rdf:type <given_class> ;
  ?v ?o.
  FILTER(?v NOT IN (rdf:type))
}
```

(b) A query to get all properties of *given class*

```
SELECT (GROUP_CONCAT (DISTINCT ?range ; SEPARATOR = ",") AS ?gc)
WHERE{
  ?s <given_property> ?o ;
  rdf:type <given_property>.
  ?o rdf:type ?range .
}order by (?o)
```

(c) A query to get range of *given* combination of *class and property*

FIG. 2. A series of SPARQL queries.

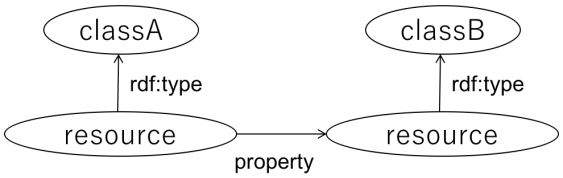


FIG. 3. Class A as subject and class B as object.

"Number of class as an object" is obtained in a similar way. Then if number of the class as a subject is larger than number of the class as an object, we define the class as main classes. Main classes and properties which are relevant to main classes are put in domain model. Third is to describe domain model as directed graph.

Experiment

We conducted an experiment to verify the validity of domain model estimated by our method. We prepared 5 LOD datasets and manually determined their correct domain model. We then estimated 5 domain models of 5 datasets using our method, followed by a comparison of the domain models made through both methods. Correct domain model is determined on the basis of published information about the metadata schema, such as text description or using RDF metadata found in the website. Correct domain model was confirmed by members of our lab. We must add that we use existing domain model, which is in graph format, if it exists. we calculated precision and recall by converting the directed graph into RDF triples. We show datasets used in experiment TABLE 1. We used 0.1% of all data found in Europeana. The purpose of this was to verify whether our method is useful when applied to a portion of metadata instances.

Results and Discussion

Table 2 shows results of the experiment. As can be seen in TABLE 2, the results from Europeana and Kyoto Kokusai Manga Museum (KKM) were below standard. In Europeana, our conclusion is

that there was an insufficient number of metadata instances. To address this, we plan to prepare a sufficient quantity of metadata instances, or establish a random sampling method for a small quantity of metadata instances. A cause of bad precision and recall in KKM is that there are some unused classes and properties in metadata instances, while still being described in published information about metadata schema. The problem of terms, which are not used in metadata instances nevertheless being described in published information, is common among many datasets. This problem needs to be discussed in the future.

TABLE 1: used datasets

Datasets name	Correct domain model	Memo
Aozorabunko LOD	Made by hand	
CiNii	Made by hand	
Europeana	Existing model	1 / 1000 of Overall
Kyoto Kokusai manga museum	Existing model	
NDLSH	Made by hand	

TABLE 2: Results of experiment.

Datasets name	precision	Recall
Aozorabunko LOD	0.85	1
CiNii	0.83	0.83
Europeana	0.07	0
Kyoto Kokusai Manga Museum	0.23	0.2
NDLSH	0.63	0.63

Relevant study

ELLIS(Gottron) is most similar study. While ELLIS provides exhaustive schema information by use of an interactive interface, our approach provides rough limited schema information by use of static image.

Conclusion

In this research, we proposed a method for estimating a domain model and conducted an experiment to verify validation of the method. We concluded that our method needs a greater number of metadata instances in order for the experiment to produce better results. A primary problem is an evaluation for validity of our method. We can't say that the purpose of domain model estimated by our method is same as purpose of existing domain model. The purpose of our domain model is to let users understand the dataset. As existing domain models are typically determined at the stage of design, the domain model often contains classes or properties which are not used in metadata instances. Therefor to verify whether coherence of LOD datasets is improved, comparing estimated domain model with existing one is not suitable. In the future, we need to verify whether the coherence of LOD datasets is improved when utilizing our estimated domain model.

References

Dublin Core singapore-framework. Retrieved February 19, 2017, from http://dublincore.org/documents/singapore-framework/ .

Aozorabunko LOD. Retrieved February 19, 2017, from http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/ .

CiNii. Retrieved February 19, 2017, from https://support.niii.ac.jp/ja/CiNii/api/api_outline .

Europeana. Retrieved February 19, 2017, from http://pro.europeana.cu/ .

Kyoto kokusai manga museum. Retrieved February 19, 2017, from http://mdlab.slis.tsukuba.ac.jp/lodc2012/kmm/ .

Web NDL Authorities. Retrieved February 19, 2017, from http://id.ndl.go.jp/information/download/ .

Tsunagu Honma, Mitsuharu Nagamori, and Shigeo Sugimoto. (2014). Extracting Description Set Profile from RDF Datasets using Metadata Instances and SPARQL Queries. Graduate School of Library Information and Media Studies University of Tsukuba. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2014, 109-118.

Thomas Gottron, Malte Knauf (2016). ELLIS: Interactive Exploration of Linked Data on the Level of Induced Schema Patterns. ESWC

Creating a Linked Data-Friendly Metadata Application Profile for Archival Description

Poster

Matienzo, Mark A.
Stanford University, U.S.A.
matienzo@stanford.edu

Roke, Elizabeth Russey
Emory University, U.S.A.
elizabeth.roke@emory.edu

Carlson, Scott
Rice University, U.S.A.
sjc5@rice.edu

Keywords: archival description; linked data; archives; Schema.org; metadata mapping

Abstract

We provide an overview of efforts to apply and extend Schema.org for archives and archival description. The authors see the application of Schema.org and extensions as a low barrier means to publish easily consumable linked data about archival resources, institutions that hold them, and contextual entities such as people and organizations responsible for their creation.

Rationale and Objectives

Schema.org has become one of the most widely recognized and adopted mechanisms for publishing structured data on the World Wide Web, and has incorporated extensions to address the needs of specialist communities (Guha, et al., 2016). It has been used with some success in cultural heritage sector through libraries and digital collections platforms using both Schema.org core types and properties, as well as SchemaBibExtend, an extension for bibliographic information (bib.schema.org, n.d.). These uses include leveraging it as a means to improve search engine rankings (Scott, 2014), to publish library staff directories (Clark and Young, 2017) and to expose linked data about collections materials (Lampron, et al., 2016). However, the adoption of Schema.org in the context of archives has been somewhat limited.

Our project focuses on identifying pragmatic methods to publish linked data about archives, archival resources, and their relationships, and to identify gaps between existing models. In our initial round of work, we are looking at applying Schema.org as the core model, and are investigating and contributing to the proposed Schema Architypes extensions (W3C Schema Architypes Community Group, 2017e). We see this as an opportunity to demonstrate the potential of Schema.org as a minimally viable mechanism for publishing linked data about archives, their collections, and the entities involved in their creation and management. In addition, this project operates in the context of a larger area of effort, focused on providing archivists, metadata professionals, and technologists hands on experience with data model and ontology development.

We have identified a small number of key objectives for this initial round of work, including developing mappings to Schema.org and associated extensions such as Schema Architypes from archival description standards for search engine optimization and general web discovery; ideally producing RDF-modeled representations of archival description directly from archives management systems, rather than from representations exported from such systems (cf. Gracy, 2015); and alignment with other related data models and application profiles.

Work Plan

Our work plan for the initial areas of investigation contains two phases. The first phase of the project, completed in April 2017, involved a survey of the landscape of related initiatives, and the identification of use cases which informed the objectives listed above. Our landscape survey focused on providing an initial review of potential models to serve as the basis for this work, including Schema.org; the Linking Lives project (Stevenson, 2012); the Bibframe Lite archives extension (Zepheira, n.d.; Zepheira and Atlas Systems, 2016); and the Europeana Data Model

(Hennicke, et al., 2011). The group chose not to evaluate the draft Records in Context Conceptual Model (International Council on Archives Expert Group on Archival Description, 2016) for this purpose given its complexity, the lack of an associated ontology (originally scheduled to be released in late 2016), and the likelihood of substantial revision. Through this discussion, we decided to continue work on investigating a Schema.org profile for archives for four reasons: its simplicity and suitability towards both providing a basic representation of entities identified in archival description; the preexisting work on the Schema Architypes extension; the need to contribute domain expertise to the W3C Schema Architypes Community Group; and the opportunity to incorporate Schema.org markup directly into archival management and discovery applications, providing a representation suitable for search engines and easier consumption by other downstream client applications, such as request management systems. The Community group developed several modeling approaches (W3C Schema Architypes Community Group 2017a, 2017b, 2017c, 2017d; see figures 1 and 2) and submitted formal proposals to the Schema.org community in September 2017 (Wallis 2017a, 2017b; see figure 2). The types and properties introduced in the proposals are strong contenders to address our use cases related to search engine optimization, improved discovery, and consumption by client applications.

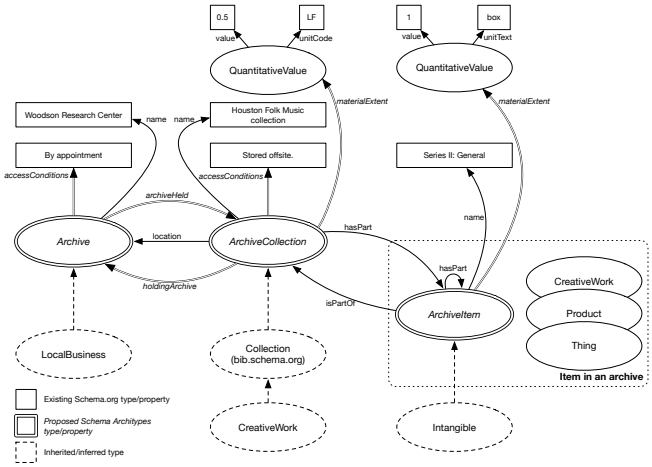


FIG. 1. Initial Schema Architypes proposal extension with extent extension. Adapted from W3C Schema Architypes Community Group 2017b, 2017d.

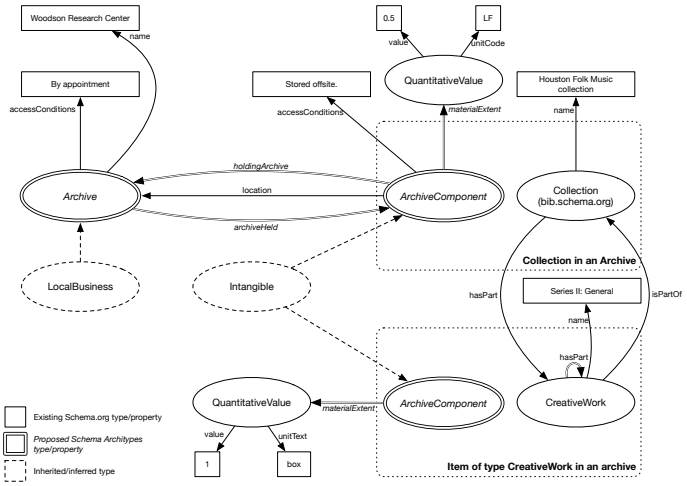


FIG. 2. Alternative Schema Architypes proposal with extent extension, submitted as Schema.org proposals (Wallis 2017a, 2017b). Adapted from W3C Schema Architypes Community Group 2017a, 2017b.

The second phase of the project, completed and pending feedback as of August 2017, was to undertake in-depth analysis of Schema.org and its associated extensions as a means to develop a profile suitable for publishing linked data for archives. Specific deliverables for this phase include identification of archival descriptive elements that should be expressed in Schema.org and undertaking a gap analysis of existing Schema.org and Schema Architypes types and properties; creating examples of Schema.org-based archival description; direct feedback and proposed revisions to the Schema Architypes Community Group; developing mappings from both content and structure standards for archival description (including ISAD(G), ISAAR-CPF, DACS, and Encoded Archival Description); and developing recommended mappings from data models of open source archives management applications such as ArchivesSpace (Matienzo and Kott, 2013) and AtoM (Artefactual Systems, 2015) to this profile. As of late May 2017, we have a completed a preliminary set of mappings from ISAD(G), ISAAR-CPF, DACS, and the ArchivesSpace and AtoM data models to Schema.org and the Architypes extensions for collection-level descriptions and information about agents and archival repositories, and have created a small number of draft examples used to verify our mappings. (See figure 3 for an example of DACS elements mapped to Schema extensions.)

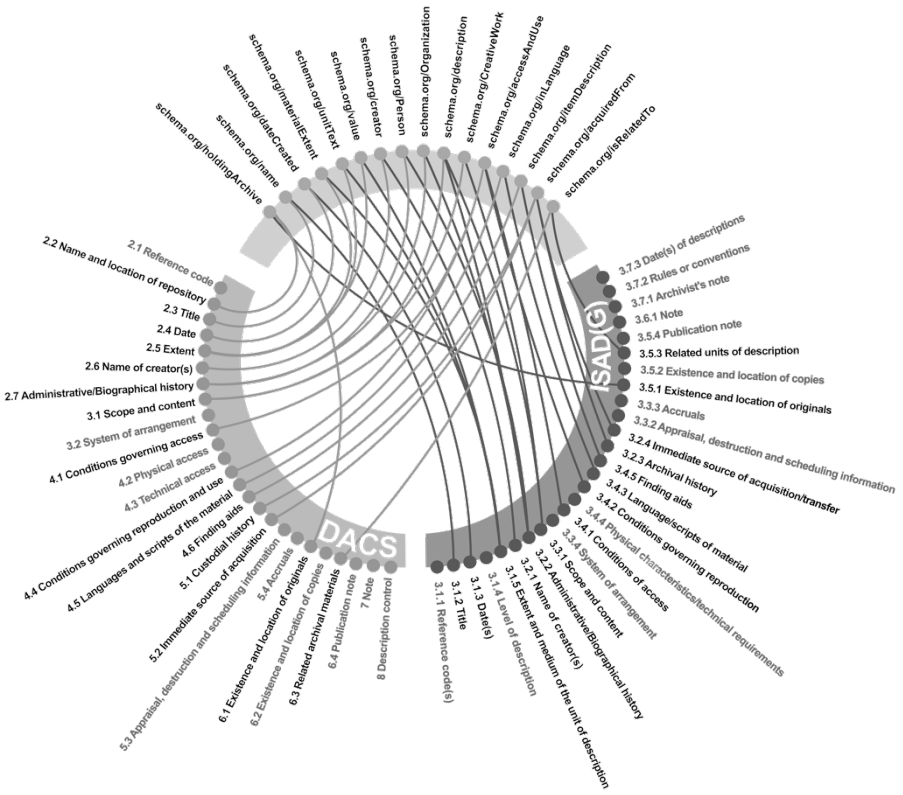


FIG. 3. Elements from *Describing Archives: A Content Standard* (bottom left) and *ISAD(G)* (bottom right) mapped to appropriate Schema.org properties. Note the *DACS* or *ISAD(G)* elements in gray, which do not yet have applicable Schema mapping.

The process of mapping existing descriptive standards and application-specific data models to Schema.org and the Schema Architypes extensions was mostly straightforward, with a few notable exceptions. The most substantive discussion within our group occurred around the desire or utility of mapping information in the description control area (e.g. *ISAD(G)* §3.7), which relates to information about the archival description itself, such as standards used, date of

descriptions, and the like. After careful consideration, we chose to not to map this information and instead emphasized representation of collection metadata over metadata about finding aids. Our group also discussed the complications of mapping the level of description for a given unit (*ISAD(G)* §3.1.4), given a lack of consistency in existing practice and data, and a discussion about its direct relevance to addressing use cases around search engine optimization and improved discovery. These conversations led to a decision to not map this data despite its perceived importance by archivists. We believe this concern may be alleviated by using *isPartOf* and *hasPart* relationship properties expressed in Schema.org to emphasize contextual relationships across levels of description within an archival collection. Reference codes (*ISAD(G)* §3.3.1) were also identified as an area for further consideration given a lack of clarity in existing archival practice. While we investigated patterns developed for SchemaBibExtend for call numbers and barcodes (W3C Schema Bib Extend Community Group, 2015), we found these patterns to be ambiguous given widely varying practices in how reference codes are assigned or used by archival repositories. Beyond these areas, suitable mappings still have yet to be identified for information usually expressed as textual notes, such as information about appraisal, accruals, or arrangement (*ISAD(G)* §3.3.2-3.3.4); physical characteristics and technical requirements related to access (*ISAD(G)* §3.4.4); and references to originals or copies of archival material (*ISAD(G)* §3.5.1-3.5.2). We expect additional feedback from archivists, metadata professionals, and other stakeholders will allow us to identify candidate mappings for these gaps, and will provide the necessary feedback to validate or refine our analysis.

Expected Benefits and Future Work

Our project provides a satisfactory proof of concept and test corpus of information about archives that will serve as a basis for fuller implementations. We believe that this will additionally allow institutions to better understand limitations in their existing descriptive data. To that end, the group is actively soliciting additional examples of archival description expressed using Schema.org and the proposed extensions (Archives and Linked Data Interest Group, 2017). Given our focus in mapping from archives management systems to a profile based on Schema.org and Schema Architypes, we see the opportunity to implement this profile directly in applications designed to support discovery of archival information, such as the public user interfaces provided by management systems like ArchivesSpace and AtoM, as well as other open source archival discovery-focused applications such as staticAid (Arnold, et al., 2017) and ArcLight (Stanford University Libraries, 2017). In addition, we expect to extend our work to undertake more in-depth investigation of and mapping to other proposed ontologies and data models for archives, with the possibility of generating extension ontologies or application profiles through further gap analysis.

Acknowledgements

Special thanks to the members of the Archives and Linked Data interest group working on this project: Scott Carlson, Mark Custer, Patrick Galligan, Dan Gillean, Gloria Gonzalez, Maggie Hughes, Mark Matienzo, Dave Mayo, Laney McGlohon, Evelyn McLellan, Katy Rawdon, and Elizabeth Russey Roke.

References

Archives and Linked Data Interest Group. (2017). Schema.org and Schema Architypes for linked archival description. Retrieved October 6, 2017 from https://archival.github.io/schema-org/

Arnold, Hillel, Kevin Clair, Luke Scott, Erin O'Meara, and Scott Carlson. (2017). staticAid. Retrieved October 6, 2017 from https://github.com/helrond/staticAid

bib.schema.org. (n.d.). Retrieved October 6, 2017 from http://bib.schema.org/

Artefactual Systems. (2015). AtoM: open source archival description software. Retrieved October 6, 2017 from https://www.accesstomemory.org/en/

Clark, Jason A. and Scott W. H. Young. (2017, April). Linked data is people: building a knowledge graph to reshape the library staff directory. Code4lib Journal, 36. Retrieved October 6, 2017 from http://journal.code4lib.org/articles/12320

Describing Archives: A Content Standard (DACS) (2nd ed.). (2015). Chicago: Society of American Archivists. Retrieved October 6, 2017 from https://www2.archivists.org/standards/DACS

Gracy, Karen. (2015). Archival description and linked data: a preliminary study of opportunities and implementation challenges. Archival Science 15, 239-294. doi:10.1007/s10502-014-9216-2

Guha, R.V., Dan Brickley, and Steve Macbeth. (2016, February). Schema.org: evolution of structured data on the web. Communications of the ACM, 59(2), 44-51. doi:10.1145/2844544

Hennicke, Steffen, Victor de Boer, Antoine Isaac, Marlies Olensky, and Jan Wielemaker. (2011). Conversion of EAD into EDM linked data. In *Proceedings of the First International Workshop on Semantic Digital Archives*, Berlin, 2011. Retrieved October 6, 2017 from http://ceur-ws.org/Vol-801/paper7.pdf

International Council on Archives Experts Group on Archival Description. (2016, September). Records in Context: a conceptual model for archival description. Consultation draft v0.1. Retrieved October 6, 2017 from http://www.ica.org/sites/default/files/RiC-CM-0.1.pdf

ISAD(G): General International Standard Archival Description (2nd ed.). (1999). Ottawa: International Council on Archives.

Lampron, Patricia, Jeff Mixter, and Myung-Ja K. Han. (2016). Challenges of mapping digital collections metadata to Schema.org: working with CONTENTdm. In E. Garoufallou, I. Subirats Coll, A. Stellato, J. Greenberg (Eds.), *Metadata and Semantics Research: MTSR 2016*. Communications in Computer and Information Science, vol 672.

Matienzo, Mark A., and Katherine Kott. ArchivesSpace: a next-generation archives management system. (2013, April). Paper presented at MW2013: Museums and the Web 2013, Portland, Oregon. Retrieved October 6, 2017 from http://mw2013.museumsandtheweb.com/paper/archivesspace-a-next-generation-archives-management-system/

Scott, Dan. (2014). Seeding structured data by default via open source library systems. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (Eds.), *The Semantic Web: Trends and Challenges: ESWC 2014*. Lecture Notes in Computer Science, vol 8465.

Stevenson, Jane. (2012). Linking Lives: creating an end-user interface for linked data. Information Standards Quarterly 24(2/3), 14-23. doi:10.3789/isqv24n2-3.2012.03

Stanford University Libraries. (2017). ArcLight. Retrieved October 6, 2017 from https://wiki.duraspace.org/display/samvera/ArcLight

Wallis, Richard. (2017a). Archives and their collections. In *Schema.org* (Github repository). Retrieved October 6, 2017 from https://github.com/schemaorg/schemaorg/issues/1758

Wallis, Richard. (2017b). MaterialExtent & CollectionSize. In *Schema.org* (Github repository). Retrieved October 6, 2017 from https://github.com/schemaorg/schemaorg/issues/1759

W3C Schema Architypes Community Group. (2017a). Alternative model proposal 1. Retrieved October 6, 2017 from https://www.w3.org/community/architypes/wiki/Alternative_1_model_proposal

W3C Schema Architypes Community Group. (2017b). Extent proposal. Retrieved October 6, 2017 from https://www.w3.org/community/architypes/wiki/Extent_proposal

W3C Schema Architypes Community Group. (2017c). Initial model. Retrieved October 6, 2017 from https://www.w3.org/community/architypes/wiki/Initial_model

W3C Schema Architypes Community Group. (2017d). Initial model proposal. Retrieved October 6, 2017 from https://www.w3.org/community/architypes/wiki/Initial_model_proposal

W3C Schema Architypes Community Group. (2017e). Schema Architypes Community Group. Retrieved October 6, 2017 from https://www.w3.org/community/architypes/

W3C Schema Bib Extend Community Group. (2015). Holdings via offer. Retrieved October 6, 2017 from https://www.w3.org/community/schemabibex/wiki/Holdings_via_Offer

Zepheira. (n.d.) Bibframe Lite + Archives. Retrieved October 6, 2017 from http://bibfra.me/view/archive/

Zepheira and Atlas Systems. (2016, February). Linked data for archives focus group report: findings from conversations with Zepheira and Atlas Systems. Retrieved October 6, 2017 from https://docs.google.com/document/d/1wEX-b_LnJJmJNKpu82r9cBqhV2sk7F9hFVaFSjfVhMY/edit

Collaborative Metadata Application Profile Development for DAMS Migration

Poster

Anne M. Washington
University of Houston, USA
awashington@uh.edu

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

Keywords: metadata; migration; application profiles; metadata schemas; digital asset management systems; CONTENTdm; Hyku

Abstract

In 2015, after an extensive review process, the University of Houston (UH) Libraries chose the open source systems Hyku (then known as the Hydra-in-a-Box project), Archivematica, and ArchivesSpace to form the Libraries' digital collections access and preservation ecosystem (Wu et al., 2016). This suite of systems, along with locally developed tools, form the Bayou City Digital Asset Management System (BCDAMS). In 2016, the BCDAMS Implementation Team began work on a multi-phase process to roll out the new systems to replace the current digital collections management system, CONTENTdm. Phase I of this process included developing fundamental models and principles as well as much of the local infrastructure and workflows (Weidner et al., 2017). Phase II of the project will involve migrating existing digital collection metadata and files to the new digital asset management system (DAMS).

This poster summarizes work done during Phase I of the project to prepare for the migration work in Phase II. This included working collaboratively to develop a Metadata Application Profile (MAP) and crosswalk for the Hyku digital collections access system, and an analysis of metadata remediation required to prepare for migration. It addresses work underway at many institutions exploring or actively migrating to a new DAMS. This poster shares the UH Libraries unique experience in preparing for the migration of UH Digital Library (UHDL) data from CONTENTdm to a new system and offers some general considerations for migrations.

To develop the MAP and crosswalk from CONTENTdm, the Descriptive Metadata Working Group (DMWG) was formed. This interdepartmental team represented the Metadata and Digitization Services Department and Special Collections. Led by the Metadata Coordinator, who is also the BCDAMS Project Manager, it included the Metadata Librarian, Metadata Unit staff members, the Coordinator of Digital Projects, the Hispanic Archivist, and the Special Collections Project Manager. It was important to have these different perspectives represented to address the metadata needs of different collections, further a shared understanding of the scope and function of the system, and gain support from stakeholders such as Special Collections curators.

Foundational migration considerations that informed the group's work were the types of content in the digital library, the purpose of the DAMS, and the technical specifications of the DAMS. As part of the DAMS evaluation process and preparation for system migration, an inventory of content types in the UHDL was completed. The Metadata Unit staff reviewed all of the digital collections individually recording the types of items it contained, e.g. single-sided images, double-sided images, documents, single-part audio, etc. In its discussions, the DMWG considered the descriptive needs of these content types, as well as content types that may be included in the digital library in the future, such as born digital content. This analysis also made clear collections' varying levels of complexity. Eight collections of low or moderate complexity were selected as test collections, used to test software and workflows.

Early in the work of both the BCDAMS Implementation Team and the DMWG, it was important to ensure a shared understanding of the purpose and scope of each component of the digital asset management and preservation ecosystem. Hyku was defined as the access system with the primary purpose of digital object discovery and access. Maintaining archival context of digital objects is especially important to Special Collections staff and users, but extensive metadata related to that context is out of scope for the access system. Instead, ArchivesSpace was determined to be the system of record for that information and a digital object's Archival Resource Key (ARK) and ArchivesSpace identifier would maintain the connection between digital objects in the access system and the physical items managed in ArchivesSpace. Once these purposes were clear, it was easier to have conversations around metadata fields and input guidelines that were appropriate for the different systems.

The technical specifications of the access system influenced the earliest MAP decisions and the work of the DMWG. At the time the DMWG was developing the Metadata Application Profile, Hyku was in early stages of development, and the built-in metadata structure was unclear. It was known that Hyku would provide a simple way for institutions to get their content to the Digital Public Library of America (DPLA), so it was assumed there would be a way to align the Hyku metadata with the DPLA Metadata Application Profile. Because of this interoperability, as well as the appropriateness of the schema for UHDL content, the team used the DPLA MAP v4 (DPLA, 2015) as the basis for the Metadata Application Profile. Given the extensible nature of Samvera Community (formerly Hydra Community) software, the team also considered additional elements not included in the DPLA MAP as long as they were elements from a linked-data ready schema such as BIBFRAME.

The team used a variety of tools in its work. The team communicated on Slack, a messaging application, between meetings. GitHub was used not only as a platform to publish the MAP code, but also to record discussions about MAP fields and input guidelines, as well as document the MAP and crosswalks. This content is openly available on the UH Libraries BCDAMS MAP GitHub repository: https://github.com/uhlibraries-digital/bcdams-map. The group also analyzed reports of existing UHDL metadata to inform decisions on the use of fields in the new MAP. These reports were created using Hunting, a locally developed Ruby gem used to access UHDL metadata through the CONTENTdm API (UH Libraries, 2016a). While there were established input guidelines in the UHDL metadata dictionary (UH Libraries, 2016b), the reports were useful in determining how the field was used in practice across different collections. The final deliverables for this group were a machine actionable MAP - the metadata element set and the input guidelines (UH Libraries, 2017a) - and a crosswalk from the existing MAP (UH Libraries, 2017b) to the new MAP.

After this work was completed, the Metadata Unit sought to understand the scope and scale of metadata remediation efforts required for migration. The Metadata Coordinator created a report for each digital collection that mapped the existing data into the new fields. The Metadata Unit staff then reviewed these reports noting where the field values did not meet the new MAP input guidelines. The most common issues recorded were: titles, subjects, dates, and fields capturing format and physical characteristics did not conform to new input guidelines. For example, previous input guidelines required unique titles, resulting in many titles ending in "Image 1" or "Image 2" to disambiguate otherwise identical titles. This is no longer a requirement in the new MAP as each resource is assigned a persistent ARK. Another new input guideline discourages the use of pre-coordinated subject headings, which were used in nearly every legacy collection. In Phase II, the Metadata Unit will begin upgrading the metadata by bringing existing data into alignment with the new MAP as well as doing authority control using the UH Libraries' local thesaurus application, Cedar (UH Libraries, 2016c).

Work has just begun on Phase II of the project which includes data migration to the new system. With its charge complete, the Descriptive Metadata Working Group has changed direction and expanded into the Data Migration Working Group. It now includes representation from the Digitization Unit to advise on file management requirements for the migration. Digital

collection curators and other stakeholders will join the group as needed to advise on specific collection concerns or other areas of development. The initial goal of this group is to determine the workflow for migration starting with the test collections. Then, after migration begins in earnest, the team will work collection by collection to address its specific file management and metadata needs. There are challenges on the horizon including technical constraints and idiosyncratic digital collection data, but the foundational work of Phase I and the commitment to a collaborative approach to migration set the stage for success.

References

DPLA. (2015). Metadata Application Profile, version 4.0. Retrieved June 12, 2017 from http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf.

Weidner, Andrew, Sean Watkins, Bethany Scott, Drew Krewer, Anne Washington, and Matthew Richardson. (2017). Outside the box: Building a digital asset management ecosystem for preservation and access. Code4Lib Journal, 36. Retrieved June 12, 2017, from http://journal.code4lib.org/articles/12342.

Wu, Annie, Santi Thompson, Rachel Vacek, Sean Watkins, and Andrew Weidner. (2016). Hitting the road towards a greater digital destination: Evaluating and testing DAMS at the University of Houston Libraries. Information Technology and Libraries, 35(2). Retrieved June 12, 2017, from https://doi.org/10.6017/ital.v35i2.9152.

UH Libraries. (2016a). hunting GitHub repository. Retrieved June 12, 2017 from https://github.com/uhlibraries-digital/hunting.

UH Libraries. (2016b). Metadata Dictionary. Retrieved June 12, 2017 from http://digital.lib.uh.edu/about/metadata.

UH Libraries. (2016c). University of Houston Libraries vocabularies. Retrieved June 12, 2017 from https://vocab.lib.uh.edu/en.html.

UH Libraries. (2017a). Bayou City DAMS metadata application profile. Retrieved June 12, 2017 from https://vocab.lib.uh.edu/bcdams-map.

UH Libraries. (2017b). UHDL crosswalk. Retrieved June 12, 2017 from https://github.com/uhlibraries-digital/bcdams-map/wiki/05-UHDL-Crosswalk.

SEPIA Project: Providing Access to Digital Image Content for the Blind and Visually Impaired.

Poster

Jennifer Sweeney
Kent State University, USA
jsween10@kent.edu

Keywords: metadata; semantic annotation; image annotation; blind and visually impaired (BVI); access; accessibility; linked open data; museums; archives; screen reader(s)

Abstract

This paper presents an introduction to the SEPIA project (SEmantic Photographic Image Annotation), which was created by Jennifer Sweeney with Blind and Visually Impaired (BVI) individuals as the designated intended user base. This project embodies a use case methodology and use-case scenario for utilizing a new data model to enhance and optimize metadata to heighten access to digital image content with screen readers.

Background

This study has shown that when a BVI user seeks to access information in a digital image collection, the typical html framework severely restricts their access to display metadata. Acknowledging that digital image content aids in enhancing our comprehension of historical and topical events, the near total lack of access for the BVI community prompted this research project. The objective is to define a methodology to create, transform, curate and enhance pre-existing collections' metadata to enable a screen reader–accessible environment. To facilitate the reconceptualization of metadata, the initial goal of the SEPIA project is to provide one use case scenario for the May 4th Collection at Kent State University. The SEPIA project seeks to create a mediator element option that would circumvent the inaccessibility of content due to access issues that are not addressed by collections software providers and typical HTML framework.

Reasoning

In the summer of 2017 the beginning stages of this project addressed two facets of BVI access issues: the first being how to mediate some of the hurdles of screen readers working with HTML in this content area, and the second being that descriptive metadata pertaining to the visual content of digital objects is not often crafted with the BVI community in mind.

Access problems often occur because website designers mistakenly assume that everyone sees and accesses a web page in the same way (ADA, 2017). The first focus of the SEPIA project is on the identification of elements (tags) within HTML that are problematic for screen readers, locating areas of code that can be transformed. The ADA presents this topic through addressing "Images Without Text Equivalents," and explains that because screen readers can only read text, they cannot interpret any digital images, so images must be annotated to provide description and context. A BVI user visiting the website would be unable to tell if the image is a photo, a logo, a map, a chart, artwork, a link to another page, or even a blank page (ADA, 2017). In Cultural Heritage Collections, it is common to have many digital image types presented on the same page, from collections logos to collections content. The solution is presented the ADA as "Add a Text Equivalent to Every Image" through alternative text embedded in the HTML using "alt" for small amounts of text and "longdesc" for long amounts. This has been the only avenue for providing context to an image object on a website, and the only opportunity to relay the same meaningful information that other users obtain by looking at the image (ADA, 2017). For example, if a BVI user is browsing a collection of Cultural Heritage objects from a civil rights collection and reads a

caption that says "Protesters holding signs" then navigates to the alternative text and is presented with the description "CivilRights0002.jpg," how is it possible to classify this content as accessible? In practice however, this action is considered in compliance because the descriptive tag is included, but it does not actually facilitate the understanding of the content. It is also important to note that as web-based digital collections shift further towards creating engaging user experiences for the sighted, the phrasing of HTML is altered through the inclusion of CSS and javascript for formatting, rendering many of the tags that screen readers depend on become hidden and create more difficulties for the BVI user.

Methodology

The project began with access to the Kent State May 4th Digital Collection with the goal in mind "that all information is available in a form that can be perceived by all users" (WCAG 2.0, 2017), and found that each of the 30 images in the University News Service photographs: Boxes 28 collection can be accessed on a dedicated page that includes an OMEKA image viewer frame embedded within the text content. This project discovered the specific access issues presented by the OMEKA CMS platform and focused on creating a collection-specific solution.

Phase 1 - The first goal of the project was to create a platform for reconceptualizing the way that descriptive metadata is written about digital objects. Collections data is typically technical rather than descriptive, so writing more effective descriptions for collections material will benefit all users, no matter of their ability. Drawing heavily Panofsky's and Barthes's writings on art theory, a full assessment of visual content was created to enable rich narrative descriptions of images, and as well as a data dictionary and term database for future data initiatives.

Phase 2 - Once the new data was created, the project shifted to testing the collection through numerous screen readers to identify specific areas of access issues and investigate where such issues existed within the HTML. After unpacking the OMEKA records and general framework of the greater Kent State site, it was found that for a screen reader, portions of the metadata associated with the images were masked by many layers of HTML. Realizing the need for a mediator element between the sea of HTML and the BVI user the SEPIA project began experimenting with javascripts and modal boxes to pull specific lines of the OMEKA record out of the body of the HTML.

Phase 3 - The final phase of the project consisted of building a mock copy of the Kent State Collection site and supplying it with a locally hosted data store that included the recoeptualized metadata. After placing the new code into the mock site, a small icon was created at the top of the page that created a mediator window linking directly to the descriptive metadata. Testing with two different screen readers in both Safari and Chrome browsers on Mac and Windows operating systems, the screen readers proved that the content was more easily accessible, provided a better user experience, and aided in the information-seeking procedure

Example of enhanced metadata entry

OMEKA XML export from site,

```
<element elementId="41">
      <name>Description</name>
      <description/>
      <elementTextContainer>
        <elementText
elementTextId="19213">
          <text>Close-up of man addressing
the crowd at the Victory Bell (W.H.O.R.E.
event, burying the Constitution)</text>
        </elementText>
      </elementTextContainer>
    </element>
```

Rewritten XML

```
<element elementId="41">
      <name>Description</name>
      <description/>
      <elementTextContainer>
        <elementText elementTextId="19213">
          <text> Black and white photograph
captured on May 1, 1970 during the
W.H.O.R.E. event when organizers were
burying a copy of the U.S. Constitution in
protest of American troops invading
Cambodia. The main subject is three male
students addressing the crowd while standing
on a small brick wall adjacent to the Victory
Bell. The Victory Bell and the crowd are not
seen in the photograph as it is taken from the
perspective of the crowd. There is one man
holding a microphone attached to a
loudspeaker up to the face of another man who
is holding a stack of papers. </text>
        </elementText>
      </elementTextContainer>
    </element>
```

FIG. 1. Phase 1

Adding the SEPIA modal box to the HTML

```
<style>
                #next-item {float: right !important;}
              </style>
              <!--<p>Sepia Test</p> <-->
              <!-- Trigger/Open The Modal -->
              <button id="myBtn">SEPIA</button>

              <!-- The Modal -->
              <div id="myModal" class="modal">

                <!-- Modal content -->
                <div class="modal-content">
                  <span class="close">&times;</span>
                  <li><a href="test.html"><b>Description</b>:Black and white photograph
captured on May 1, 1970 during the W.H.O.R.E. event when organizers were burying a copy
of the U.S. Constitution in protest of American troops invading Cambodia. The main subject is
three male students addressing the crowd while standing on a small brick wall adjacent to the
Victory Bell. The Victory Bell and the crowd are not seen in the photograph as it is taken from
the perspective of the crowd. There is one man holding a microphone attached to a loudspeaker
up to the face of another man who is holding a stack of papers.  </a></li>
                  <br>
```

FIG. 2. Phase 2

Limitations

This project was presented as a MLIS research paper and project with a time construct of eight weeks. In the true nature of research and development, every week presented a new hurdle and solution to test and consider. The SEPIA use case was created by mirroring data from the Kent State collection, rather than building upon it. Due to the inability to parse actual metadata records through the code to the HTML, this use case does not present a final project, but rather a working model.

Future

The project is currently in the testing and application state. With a strong belief in the data model and the utilization of HTML alterations, the extent of the capabilities that the SEPIA tool presents is unknown.

Future goals are to identify, research and work with collections that want to utilize tools in the digital space that allow for regional markup and annotation of visual image content. This project imagines a future when even the most dynamic web content can be translated to the BVI user group, providing comparable access to the content. This research has shown that with a mindful and conceptual approach to this problem of access, the Library Archive and Museum community can not only create better information resources on the web, but also enable a path for reconceptualizing an inclusive user experience. When the SEPIA project is poised for deployment to the public, it will be a remarkable tool to aid in ending the designation that BVI users are the "second-class citizens of the information society" (Jaeger, 2008).

Where can this go in the long run? Cultural Heritage Institutions can utilize this model across a range of collections and access platforms. From including the SEPIA markup on websites and smartphone apps, to adding similar narrative elements to audio tours of physical exhibitions, the data model can be applied for virtually all avenues of access, opening collections to a previously underserved audience.

References

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

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

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

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

ADA (2017). ADA Tool Kit: Website Accessibility Under Title II of the ADA. ADA Tool Kit: Website Accessibility Under Title II of the ADA. Ada.gov. Retrieved from https://www.ada.gov/pcatoolkit/chap5toolkit.htm

Jaeger, P. (2008). Developing Online Community Accessibility Guidelines for Persons With Disabilities and Older Adults. Journal Of Disability Policy Studies, 20(1), 55-63. http://dx.doi.org/10.1177/1044207308325997

WCAG 2.1. (2017). Working Draft for Success Criteria. Web Content Accessibility GuidelinesW3.org. Retrieved from https://www.w3.org/TR/WCAG21/

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

ISSN: 1939-1366 (Online)