

# ISQ INFORMATION STANDARDS QUARTERLY

SPRING/SUMMER 2012 | VOL 24 | ISSUE 2/3 | ISSN 1041-0031

TOPIC: LINKED DATA FOR LIBRARIES, ARCHIVES, AND MUSEUMS

LINKED DATA VOCABULARY  
MANAGEMENT

EUROPEANA MOVING TO  
LINKED OPEN DATA

LINKING LIVES  
END-USER INTERFACE

JOINING THE LINKED  
DATA CLOUD

OCLC'S LINKED DATA  
INITIATIVE

LC'S BIBLIOGRAPHIC  
FRAMEWORK INITIATIVE

**NISO**

How the information world  
CONNECTS

# 2012 NISO EDUCATIONAL EVENTS

[www.niso.org/news/events](http://www.niso.org/news/events) 

## Webinar Subscription Package Discounts

Buy all 14 for the price of 7!  
Buy 4 and get 3 free.

Discounts available for NISO members, students, NASIG members, and DCMI members for specified events. (See webpage for details.)

## NISO Open Teleconferences

Join us each month for NISO's Open Teleconferences—an ongoing series of calls held on the second Monday of each month as a way to keep the community informed of NISO's activities. The calls also provide an opportunity for you to give feedback to NISO on our activities or make suggestions about new activities we should be engaging in. **The call is free and anyone is welcome to participate in the conversation.** All calls are held from 3:00 - 4:00 p.m. Eastern time.

## AUGUST

- 8 Content on the Go: Mobile Access to E-Resources (*NISO Webinar*)
- 22 Metadata for Managing Scientific Research Data (*NISO/DCMI Joint Webinar*)

## SEPTEMBER

- 10 KBART (Knowledge Base And Related Tools) Update (*NISO Open Teleconference*)
- 12 Understanding Critical Elements of E-books: The Social Reading Experience of Sharing Bookmarks and Annotations (*NISO Webinar*)
- 24 Tracking it Back to the Source: Managing and Citing Research Data (*NISO Forum, Denver, CO*)
- 26 Discovery and Delivery: Innovations and Challenges (*NISO Webinar*)

## OCTOBER

- 10 MARC and FRBR: Friends or Foes? (*NISO Webinar*)
- 15 NISO E-Book Special Interest Group Update (*NISO Open Teleconference*)

- 18–19 The E-book Renaissance, Part II: Challenges and Opportunities (*NISO Forum, Boston, MA*)
- 24 Embedding Linked Data Invisibly into Webpages: Strategies and Workflows for Publishing with RDFa (*NISO/DCMI Joint Webinar*)

## NOVEMBER

- 14 Beyond Publish or Perish: Alternative Metrics for Scholarship (*NISO Webinar*)
- 17 Resource Synchronization Update (*NISO Open Teleconference*)

## DECEMBER

- 10 NCIP (NISO Circulation Interchange Protocol) Update (*NISO Open Teleconference*)
- 12 Connecting the Dots: Constellations in the Linked Data Universe (*NISO Webinar*)

# ISQ

SPRING/SUMMER 2012 | VOL 24 | ISSUE 2/3 | ISSN 1041-0031

## NISO

How the information world  
CONNECTS

INFORMATION STANDARDS QUARTERLY (ISQ) is a publication by the National Information Standards Organization (NISO). ISQ is NISO's print and electronic magazine for communicating standards-based technology and best practices in library, publishing, and information technology, particularly where these three areas overlap. ISQ reports on the progress of active developments and also on implementations, case studies, and best practices that show potentially replicable efforts.

NISO MANAGING DIRECTOR and ISQ PUBLISHER | Todd Carpenter

ISQ MANAGING EDITOR | Cynthia Hodgson

NISO ASSOCIATE DIRECTOR FOR PROGRAMS | Nettie Lagace

NISO BUSINESS DEVELOPMENT & OPERATIONS MANAGER | Victoria Kinnear

DESIGN | B. Creative Group, Inc.

### ISQ Board Members

Marshall Breeding, *Vanderbilt University*

Priscilla Caplan, *Florida Center for Library Automation*

Helen Henderson, *Information Power Ltd.*

Peter Murray, *LYRASIS*

Andrew Pace, *OCLC*

Kristen Ratan, *Public Library of Science (PLoS)*

Kate Wittenberg, *Ithaka*

### 2012 AD RATES

	1 ISSUE	2-3 ISSUES	4 ISSUES
Full page (8.5" x 11")	\$375	\$350	\$325
Half page (4.25" x 5.5")	\$250	\$225	\$200
Back cover (8.5" x 11")	\$700	\$600	\$550
Inside Cover Front (8.5" x 11")	\$500	\$450	\$400
Inside Cover Back (8.5" x 11")	\$500	\$450	\$400

For more information on advertising, visit [www.niso.org/publications/isq](http://www.niso.org/publications/isq)

©2012 National Information Standards Organization.

REUSE: For permission to photocopy or use material electronically from Information Standards Quarterly, ISSN 1041-0031, please access [www.copyright.com](http://www.copyright.com) or contact Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users.

## CONTENTS

2 From the Guest Content Editor

### FE FEATURE 4

4 Linked Data Vocabulary Management: Infrastructure Support, Data Integration, and Interoperability

### IP IN PRACTICE 14

- 14 Linking Lives: Creating an End-User Interface Using Linked Data
- 24 Joining the Linked Data Cloud in a Cost-Effective Manner
- 29 OCLC's Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web
- 34 Europeana: Moving to Linked Open Data

### OP OPINION 41

41 LODLAM State of Affairs

### CR CONFERENCE REPORT 43

43 Report on the Linked Ancient World Data Institute

### SP SPOTLIGHT 46

46 LC's Bibliographic Framework Initiative and the Attractiveness of Linked Data

### NR NISO REPORTS 51

### NW NOTEWORTHY 53

### SD STANDARDS IN DEVELOPMENT 56



Corey  
Harper

## FROM THE GUEST CONTENT EDITOR

---

I'm incredibly excited about this issue of *ISQ*, focused on the current state of the linked data movement within the cultural heritage sector. In 2006, when the World Wide Web consortium re-framed some of the Semantic Web concepts under the label "Linked Open Data," the underlying concepts began to gain significantly more traction—a trend which the library community rapidly became involved in. Initial forays into linked library data focused on publishing authority data using an emerging standard called SKOS (Simple Knowledge Organization System), though in recent years we've seen increased publication of linked bibliographic data alongside those authorities, and the scope of publication efforts has grown beyond the borders of libraries.

Development and change in this area has been rapidly increasing, and this issue has more of an *In Practice* project report focus than is usually the case. This is particularly exciting, as it gives a broad overview on the scope and breadth of developments happening in the world of LOD-LAM, or Linked Open Data for Libraries, Archives, and Museums. **Jon Voss** provides an opinion piece describing the LOD-LAM movement, its origins, and next steps toward planning for a global Web of interoperable cultural heritage information. From his description of the state of affairs, it's increasingly clear that LOD-LAM is gaining prominence in the archives and museum communities as well as in libraries.

Two of the *In Practice* articles come from European authors, further underscoring the international flavor of LOD-LAM. First, **Jane Stevenson** provides a project report on the Linking Lives project, a follow on from the JISC-funded Linked Open COPAC and Archives Hub project, or LOCAH. LOCAH was an initiative to publish extensive linked datasets derived from archival data, primarily finding aids, from all over the UK. Linking Lives, in turn, is one of the first LAM-based initiatives to design an end-user interface to a large collection of linked data in the library and archives domain. The article summarizes the progress, identifies limitations and challenges, and maps a path forward for the effective display and interface design for Linked Archival Data.

Later in the issue is a summary of the Europeana Project by project participants **Antoine Isaac**, **Robina Clayphan**, and **Bernhard Haslhofer**. Europeana is an ambitious EU-funded

project to improve access to European cultural heritage materials on the Web. Like Linking Lives, it is not explicitly a Linked Library Data project, but rather a massive effort to harmonize data from libraries, archives, and museums throughout Europe into a common structure and data model. The linked data components of Europeana feed into other aspects of the project, and this article is an excellent overview of the initiative as well as a roadmap for the next steps Europeana will undertake.

Continuing in the vein of cultural heritage metadata, there is a conference report from **Thomas Elliott**, **Sebastian Heath**, and **John Muccigrosso** on the Linked Ancient World Data Institute (LAWDI), a three-day workshop held in May-June 2012 bringing together 50+ researchers and digital library developers from the fields of classics, museum studies, archeology, and numismatics. This gathering was one of the first groups of academics who are *not* information professionals running an intensive workshop into the impact that linked data technology can have—and is already having—on their disciplines.

As the Linked Data Web continues to expand, significant challenges remain around integrating such diverse data sources. As the variance of the data becomes increasingly clear, there is an emerging need for an infrastructure to manage the diverse vocabularies used throughout the Web-wide network of distributed metadata. The feature article, by **Gordon Dunsire**, **Diane Hillmann**, **Jon Phipps**, and myself, discusses these infrastructure needs and describes

As the variance of the data becomes increasingly clear, there is an emerging need for an infrastructure to manage the diverse vocabularies used throughout the Web-wide network of distributed metadata.



a number of initiatives and techniques for managing diverse, heterogeneous metadata sets.

*Joining the Linked Data Cloud in a Cost-Effective Manner*, by Seth Van Hooland, Ruben Verborgh, and Rik Van De Walle, is less a project report than a practicum. The article introduces the idea of Interactive Data Transformation (IDT) tools, such as Google Refine, that provide desktop access to the linking, reconciliation, remediation, and metadata management functionality that is one of Linked Data's core strengths. Building on a case study using data from the Cooper-Hewitt National Design Museum, the article explains in detail how metadata practitioners looking to develop practical experience in LOD-LAM can begin using these IDTs to make iterative improvements to legacy metadata and begin linking it up with the myriad other data sources beginning to emerge in the cultural heritage sector.

The most library-focused articles come from two of the bigger library organizations in North America: OCLC and the Library of Congress. Ted Fons, Jeff Penka, and Richard Wallis provide an overview of OCLC's initial forays into linked data technology, including a discussion of the Virtual International Authority File and of preliminary efforts to put schema.org-based RDFa (Resource Description Framework in attributes) data into WorldCat.org pages for the millions of books and other resources in WorldCat. Kevin Ford of the Library of Congress presents an overview of LC's Bibliographic Framework Transition Initiative. The initiative, announced in October of 2011, is the initial phase in an effort to evolve and

eventually replace the MARC format with a framework that better prepares library data for inclusion in the Semantic Web. The article describes how initial modeling and the development of prototype specifications and tools are building on Linked Data principles to provide a set of metadata specifications that will enable descriptions of library resources to be more effectively integrated into the Web.

I truly hope that you enjoy these articles as much as I did, and that you come away from this issue with a more practical sense of what Linked Data can do for you, your institutions, and your resources as well as some ideas about how you can begin to implement in this space. This issue will be a success if it inspires additional projects like those discussed and helps practitioners have a better sense of what steps to take to join the LOD-LAM movement.

doi: 10.3789/isqv24n2.2012.01

**Corey A Harper** | *Metadata Services Librarian,  
New York University*

# LINKED DATA VOCABULARY MANAGEMENT:

## INFRASTRUCTURE SUPPORT, DATA INTEGRATION, AND INTEROPERABILITY

GORDON DUNSIRE, COREY HARPER, DIANE HILLMANN, AND JON PHIPPS

**Recently there has been a shift in popular approaches to large-scale metadata management and interoperability. Approaches rooted in Semantic Web technologies, particularly in the Resource Description Framework (RDF) and related data modeling efforts, are gaining favor and popularity.**

In the library community, this trend has accelerated since the World Wide Web Consortium (W3C) re-framed many of the Semantic Web's enabling technologies in terms of Linked Open Data (LOD)—a lightweight practice of using web-friendly identifiers, explicit domain models, and related ontologies to design graph-based metadata. Since that shift, the library metadata community has become an increasingly major contributor to the “global graph” of linked data. The emergence of linked data for libraries began with the Library of Congress publication of LCSH (Library of Congress Subject Headings) in SKOS (Simple Knowledge Organization System) and the Swedish National Library's publication of the LIBRIS

Union Catalog as linked data. Since then, major publishing efforts have come from the German and French national libraries, the British Library, and initiatives like Europeana, which include museum and archival data as well as data from libraries. Already, the Summer of 2012 has seen OCLC launch major linked data initiatives and the Library of Congress begin work on a Bibliographic Framework Transition Initiative based on Linked Data.

As more and more RDF-based metadata become available, a lack of established best practices for vocabulary development and management in a Semantic Web world is leading to a certain level of vocabulary chaos. The situation is aggravated by a dearth of tools for discovering and selecting existing vocabularies. This “embarrassment of riches” could be viewed as troubling proliferation or as welcome activity expanding the availability of viable approaches to description. Either way, strategies for vocabulary publishing, discovery, evaluation, and mapping have the potential to change the conversation significantly.

For the purpose of this article, “vocabulary” refers to metadata element set vocabularies (ontologies): collections of classes and properties used to describe resources in a

# VOL 24



particular domain. While many of the infrastructure components are also relevant to the management of “value vocabularies” (also called controlled vocabularies), the examples herein will be about metadata element sets. Such clarification is necessary to establish basic contexts for data expressed in the one-size-fits-all simplicity of RDF. Metadata element sets and value vocabularies, along with datasets, are contexts recently defined and scoped for archive, library, and museum linked data.<sup>[1]</sup>

## Metadata Registries

Until recently, vocabularies were considered to be tied tightly to particular domains and applications. In the library world, most vocabulary development was in the context of MARC 21, and similar development trajectories occurred within other domains of practice.<sup>[2][3][4]</sup> The first public glimmer of a less siloed approach appeared in 2000, when Heery and Patel published their seminal article on Application Profiles, a notion taken up with enthusiasm by the DC (Dublin Core) Community.<sup>[5]</sup> The idea that vocabularies could be “mixed and matched” to improve

CONTINUED »

CONTINUED »

both usefulness and interoperability was a potent one, and from that idea grew greater interest in what might be “out there” that could be reused without additional vocabulary proliferation, or the overhead of vocabulary development by every project or domain.

Even before that article, as early as 1999, metadata practitioners had begun to experiment with the idea of Application Profiles. For those innovators, the need for an infrastructure to manage discovery of and documentation for the various schemas from which terms are drawn became very clear. Early examples of work in this area include the UKOLN DESIRE Metadata Registry,<sup>[6]</sup> the European Commission funded Schemas Project, and its successor CORES.<sup>[7]</sup>

These tools became known as registries, and in 2002, the Dublin Core Metadata Initiative (DCMI) launched its own Metadata Registry.<sup>[8]</sup> According to Heery and Wagner (the DCMI registry’s initial developers):

*Metadata schema registries are, in effect, databases of schemas that can trace an historical line back to shared data dictionaries and the registration process encouraged by the ISO/IEC 11179 community.<sup>[9]</sup>*

This work has inspired a number of other registries, including the Open Metadata Registry (OMR);<sup>[10]</sup> the current version of the DCMI Registry, which has provided the basis for a national Japanese Metadata Infrastructure Registry;<sup>[11]</sup> and the JISC Information Environment Metadata Schema Registry.<sup>[12]</sup>

The OMR, among the most active of this group currently, began as the NSDL Registry, a National Science Foundation-funded project within the U.S. National Digital Library program. It was built as a free, open service and among its most important functions is the ability to provide detailed versioning of changes at every level. It has been used extensively in the library community, now hosting the vocabularies of RDA (Resource Description and Access), ISBD (International Standard Bibliographic Description) and the FR family of models (Functional Requirements for Bibliographic Records/Authority Data/Subject Authority Data) developed by IFLA (International Federation of Library Associations and Institutions), and the experimental version of MARC 21 in RDF discussed below. The OMR is now engaged in a significant redevelopment effort, focused on vocabulary mapping.

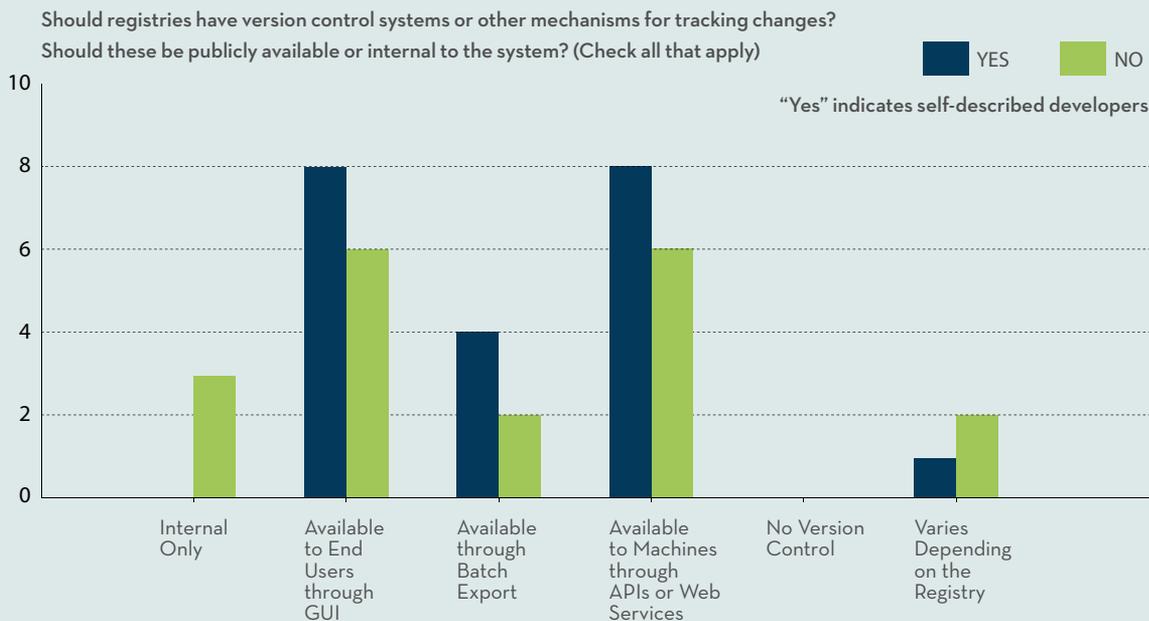


Figure 1: Version control from users

The DCMI Registry Community, established in 1999, became a central place for the discussion of the development, management, and functional requirements for metadata registries. In 2009, UKOLN, working with the DCMI Registry Community, produced a survey of Metadata Registry users and owners to identify current practice of the systems and functional requirements for vocabulary management and inter-registry interoperability. The survey, still unpublished, was completed by 12 registry owners, including most of the major active registries above, 10 self-identified application developers looking to programmatically consume registry content, and a number of other end users—to total 35 respondents.

Discrepancies between end users' needs and system functionality were seen in responses relating to types of content registered, services provided, and the data formats and methodologies used for access to content.

The chart in Figure 1, with application developers marked in dark blue and labeled "yes," shows a clear desire for machine-readable, API-based access to version history. Contrasted with Figure 2, showing that over half of the registries had no version control or did not expose that information to users, the discrepancy between the needs of registry users and the state of registry software development is evident.

The results showed that the focus of registries was becoming less about discovery of relevant vocabulary terms for mixing and matching, and more about infrastructure for managing those vocabularies, vocabulary version control, and mapping between vocabularies.

Bill de hÓra, in a 2007 blog post, stated the issues succinctly:

*There are two schools of thought on vocabulary design. The first says you should always reuse terms from existing vocabularies if you have them. The second says you should always create your own terms when given the chance.*

*The problem with the first is your [sic] are beholden to someone else's sensibilities should they change the meaning of terms from under you (if you think the meaning of terms are fixed, there are safer games for you to play than vocabulary design). The problem with the second is term proliferation, which leads to a requirement for data integration between systems (if you think defining the meaning of terms is not coveted, there are again safer games for you to play than vocabulary design).*

*What's good about the first approach is macroscopic – there are less terms on the whole. What's good about the second approach is microscopic – terms have local stability and coherency. Both of these approaches are wrong insofar as neither represents a complete solution. They also transcend technology issues, such as arguments over RDF versus XML. And at differing rates, they will produce a need to integrate vocabularies.<sup>[13]</sup>*

CONTINUED »

Does your registry have a version control system or other mechanism for tracking changes? Are various versions publicly available or internal to the system? (Check all that apply)

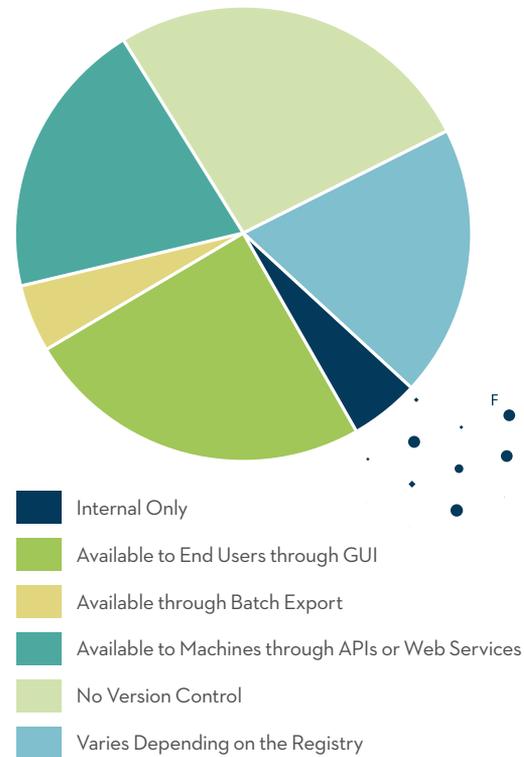
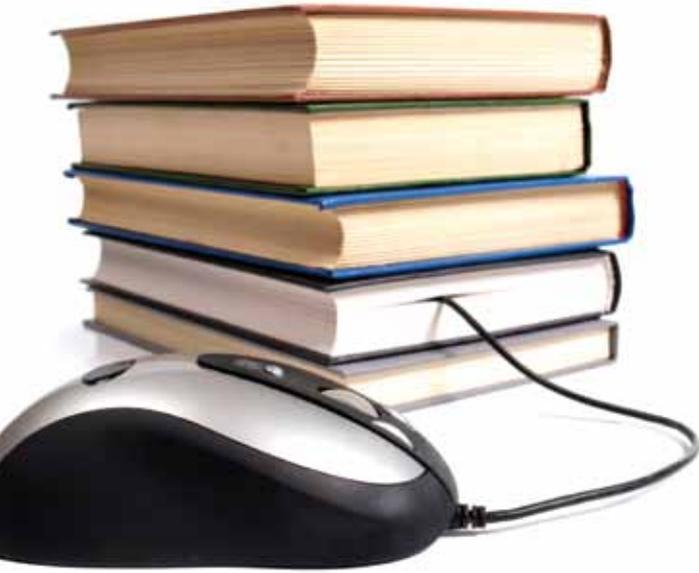


Figure 2: Version Control availability in surveyed registries



CONTINUED »

### Bibliographic Standards Communities

IFLA and JSC/COP (Joint Steering Committee for Development of RDA and Co-Publishers) are using the OMR to develop and administer RDF namespaces representing de-facto international bibliographic standards. These include the FR family, ISBD, and RDA. While technical advice and support for all of these namespaces has been provided by a small team, which includes three of the authors of this paper, the development of each set of namespaces has been largely autonomous between the standards' management infrastructure. This has identified a range of management issues to be considered.

RDA was the first of these standards to use a registry, to meet the goals of the DCMI/RDA Task Group. The development of element sets and value vocabularies for RDA<sup>[14]</sup> has taken place in an open environment, with benefits for maintainers and consumers. Version control has allowed the long development path to be monitored by external applications. The RDA namespace was created in 2008; as of July 2012 the element sets and many of the value vocabularies remain in a mutable state. Yet the visibility of status and development history has allowed experimental applications—such as those discussed below—to use RDA classes and properties in appropriate contexts. Access control allows multiple agents to work at their own pace and to develop flexible agendas for tasks such as language translations and synchronization with other documentation. Progress of, and feedback on, such work is easily monitored by colleagues and other interested parties.

The development of the RDA namespace immediately stimulated the IFLA communities to consider the potential use of their own standards in the Semantic Web, as RDA is based on the FR family. The FR element sets have followed the same development sequence as the standards, and the semantic analysis involved is informing a current process of consolidation into a single model. ISBD is developing a DC Application Profile to state requirements for a well-formed ISBD record, including mandatory and repeatable status of elements, aggregations of elements into higher-level statements, and sources of value vocabularies.<sup>[15]</sup> IFLA is also considering best practices for the translation of its element sets and value vocabularies, as it operates in a multilingual environment and recognizes seven official languages for its activities. Parts of the ISBD and FR family namespaces have been translated from English into Spanish and Croatian; translations of the underlying documentation are available in multiple languages, which might eventually be applied to the namespaces.

Reuse of RDA elements was rejected because the natural flow is to refine the application from the model. In turn, ISBD did not reuse FR elements because there was, and remains, no complete agreement on the semantic relationship between the two standards. A discussion on unconstrained namespaces for mapping between IFLA and other community metadata element sets is emerging, stimulated by work on alignment of ISBD and RDA elements to improve interoperability.<sup>[16]</sup>

This formalized and more comprehensive approach to bibliographic data is a marked contrast to earlier efforts to reuse more domain-neutral vocabularies—Dublin Core, Bibliographic Ontology (BIBO), Friend of a Friend (FOAF)—in many of the European national libraries' efforts to publish RDF representations of catalog data. Though early efforts at publishing linked library data varied in the complexity of their data model, all relied heavily on reuse of vocabularies already in wide use on the Web. Some, such as LIBRIS's trailblazing efforts, the British Library, and Cambridge University, applied existing vocabularies like BIBO and FOAF. Such projects often feature simple modeling of a few FRBR classes; associated entities representing agency, such as authorship and publication; and other entities representing aboutness, including people, places, time-periods, and topics. Others, such as the British Library's efforts, were heavily specified, with classes for information related to series, subjects, publication events, and agents.<sup>[17]</sup> The German National Library reused DC, FOAF and SKOS along with the RDA Vocabularies described above.

The Cambridge Open METadata (COMET) project, in particular, set another powerful precedent toward best practice by making all of their conversion utilities, tools, code and processes available under an open source license.<sup>[18]</sup> There is a tremendous amount of value to all of these approaches. Both the comprehensive efforts to model the rich depth of MARC 21, RDA, and ISBD and the more selective exposure of key information from that data using more common web vocabularies are important aspects of current experimentation in linked bibliographic data.

This is evidence, indeed, of the shifting balances of the macroscopic and microscopic approaches discussed by de hÓra. This has set the stage for a shift of focus in registries to the management of maps and mappings, as well as application profiles.

## The Case for Mapping

The mapping of a semantic relationship between an RDF property with another RDF property or class can be associated with an inference rule that enables the processing of data expressed using the origin property. Processing results in the generation of a new RDF statement that can be used in the environment of the target property or class. Best practice results

in many bibliographic schema attributes and relationships being expressed as RDF properties that can be included in a map (sets of mappings) as an RDF graph or ontology.

Figure 3 shows an RDF graph that maps properties with overlapping semantics for the concept “extent of a bibliographic resource.” The properties are taken from the namespaces of Bibliographic Ontology (bibo), Dublin Core terms (dct), FRBR entity-relationship model (frbrer), ISBD, MARC 21, RDA, and a proposed community-shared high-level “commons.” All links in the graph are the RDF Schema property *rdfs:subPropertyOf*, indicating a broadening of meaning in the direction of the arrow. Data using any of these namespace properties can be propagated in that direction, losing detail but preserving coherency in a “dumb down” process that provides interoperability from local to global levels.

Similar RDF graphs can be constructed for value vocabularies using the SKOS property *skos:broader*. It is a trivial technical task to incorporate vocabularies into such maps, although the information and expertise required to determine the target of each mapping should not be underestimated.

Figure 4 shows a suggested map for a single property from the *info* vocabulary<sup>[19]</sup> and equivalent properties in the *oclc:library*, ISBD, and RDA (free) vocabularies showing

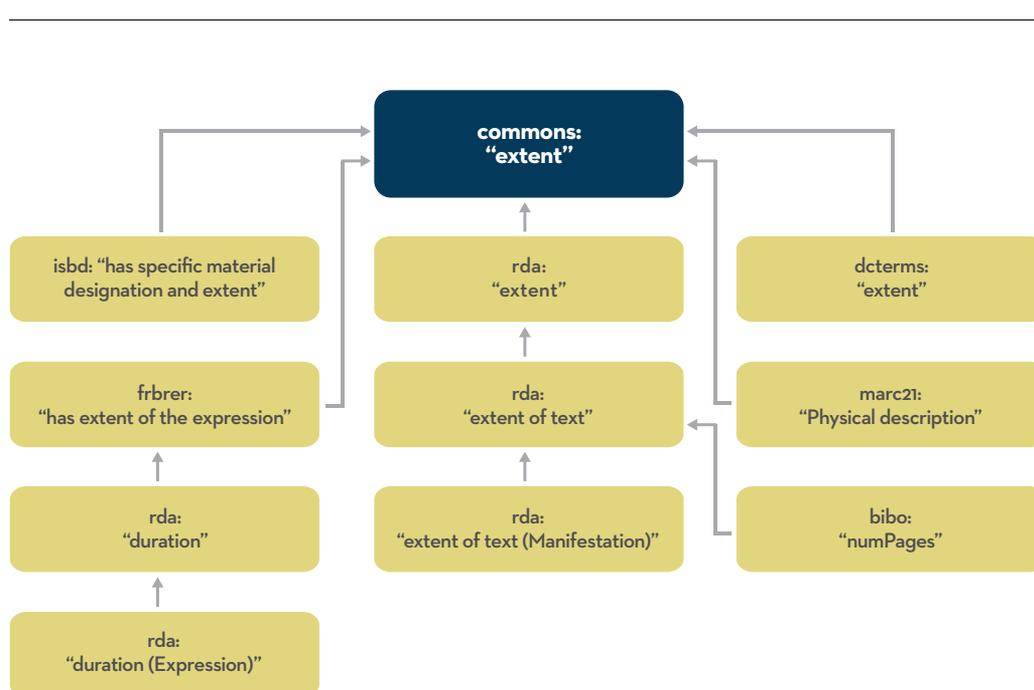


Figure 3: RDF graph of ontology for Extent

Best practice results in many bibliographic schema attributes and relationships being expressed as RDF properties that can be included in a map (sets of mappings) as an RDF graph or ontology.

Domain	Schema	Range
	marc21	
	marc21rdf:2XX/M26O_a owl:equivalentProperty	
	library	schema:Name
schema:CreativeWork	schema.org	schema:Place
	marc21rdf:2XX/M26O_a owl:equivalentProperty	
isbd:Resource	ISBD	
	marc21rdf:2XX/M26O_a owl:equivalentProperty	
frbrer:Manifestation	FRBRer	
	frbrer:Manifestation-owl:equivalentProperty	
rda:Manifestation	RDA	
	marc21rdf:2XX/M26O_a owl:equivalentProperty	
	RDA (free)	

Figure 4: Possible map for a MARC 21 property to equivalent properties in other namespaces

#### CONTINUED »

the domain and range of each. The MARC 21 vocabulary is intended to provide a completely lossless semantic mapping from MARC 21 to RDF. The URIs for each individual property have a consistent construction of *[tag][indicator 1][indicator 2][subfield]* and are designed to be programmatically constructed in order to support efficient machine-transcription. The vocabulary is specifically designed to support mapping to related bibliographic vocabularies such as ISBD, FRBRer, and RDA as well as ongoing progressive enhancement.

Note that “natural” mappings to FRBRer and RDA in this map have been removed because of the incorrect inference that the resource is a “Manifestation”. The application of multiple inference rules from a complex graph can result in semantic incoherence.

Figure 5 shows a pseudo-RDF representation of the additional metadata entailed (inferred) by the use of a single “Place of Publication” property describing an OCLC bibliographic resource and the multiple inference of its “type”, using the map in Figure 4. Note the refinement and increased accuracy of the description of “Place” provided by the *oclc:library* mapping to the original MARC 21 property. An added Google Maps URI for the actual location provides an additional enhancement.

Many expressions of MARC 21 in RDF have made the natural decision to optimize and harmonize the mapping from the necessarily complex MARC 21 syntax, with its need to express values as literal strings, to a more resource-oriented RDF, focusing on simpler descriptions of related resources as first-class entities in their own right. This is the approach taken by the British Library, LIBRIS, and other projects described earlier. While there is significant value in this optimization, there is much to be gained by also providing the original values mapped to their direct RDF equivalent. Figure 5 illustrates the value of a detailed expression of the complete MARC 21 semantics in the *marc21rdf:info* vocabulary:<sup>[19]</sup> bidirectional semantic equivalencies and subclasses can be expressed based on simple low-level mappings between semantically equivalent properties. As this example shows, by mapping at the lowest lexical level between vocabularies designed and maintained by different communities of practice, an enhancement to one can easily become an enhancement to all. Figure 5 also shows the potential for unnecessary and perhaps inaccurate entailments caused by the assignment of a too-restrictive domain. The RDA (free) vocabulary is a domain-free version of the more restrictive RDA vocabularies that was created to be used to minimize these inaccuracies when necessary.

## The Role of DCMI

During the keynote for the Dublin Core 2010 meeting in Pittsburgh, Michael Bergman prompted a change in the conversation for many members of the registry community.<sup>[20]</sup> Though registries were deemed still important, the focus shifted to their part in the general infrastructure for the management of vocabularies. Bergman's main point was to highlight an opportunity for the DCMI: given the fact that vocabulary proliferation was showing no signs of abating, he saw an emerging need for vocabulary alignment, co-referencing, and interoperability. This focus on "alignment" can be seen as somewhat analogous to the established practice of developing crosswalks between record-based (usually XML) metadata structures. Vocabulary alignment, in contrast, identifies equivalencies and other kinds of relationships between individual metadata elements to help enable the application of those properties outside the context of their source vocabularies.

However, as the notion of an open linked data environment expands, the situation we're facing is much more complex than it looks initially. As Dunsire, et al. note:

*The meaning of "mapping" changes radically on moving from a database and record based approach to an open, multi-domain, global, shared environment based on linked data technologies — where anybody can say anything about any topic, validity constraints are not acknowledged, a nearly infinite number of properties can be defined to describe an infinite number of entities, and authority is multi-dimensional and often ephemeral. The classic approach to such apparent chaos is to attempt increased control, increased filtering, increased restrictions, and limited access. This approach hinders appreciation of the broad diversity of perspective that comes with a world of open data.<sup>[21]</sup>*

Following up on the DC-2010 conversations sparked by Bergman, DCMI held a special pre-conference session at DC-2011 in The Hague<sup>[22]</sup> to identify the vocabulary management and alignment issues bedeviling the implementer communities associated with DCMI and see where DCMI could support efforts to come to grips with these issues. The result was the chartering of the DCMI Vocabulary Management Community<sup>[23]</sup> charged with identifying issues of best practice and intelligent implementation that could lead to better interoperability and harmonization across institutions, projects, and language communities.

The issues surfaced in the discussion at that session revolved around the practical problems of finding, evaluating, and using vocabularies. A strong thread of concern about

vocabulary quality and preservation underpinned the entire session—and has continued. The session conversations were intensely practical, and the questions that arose in them continue to reverberate within the Community as the group sets priorities and begins a more virtual stage of activity. The three focus areas at this point are planning for best practice guidelines around vocabulary evaluation, selection, and reuse; examining more closely the issues around vocabulary sustainability and preservation (including discussion of possible roles for DCMI); and the development of a set of best practices for principled extension of vocabularies.

A common interest in multi-lingual vocabularies also surfaced at the meeting, and conversations about available standards and tools for developing and managing vocabularies in many languages provided evidence of a strong interest in these issues. Though not surprising in an international group, this focus area will continue to be on the radar of the Vocabulary Management Community.

Significant contributions to those conversations in The Hague were made by Bernard Vatant of the Linked Open Vocabularies (LOV) Project.<sup>[24]</sup> Bernard and his team have been collecting information on extant property vocabularies

CONTINUED »

<http://www.worldcat.org/title/linked-data-evolving-the-web-into-a-global-data-space/oclc704257552>

**a schema:CreativeWork**

**a isbd: C2001 (Resource)**

**marc21rdf:2XX/M260\_a** "San Rafael, Calif.

(1537 Fourth Street..."

**isbd: P1016** "San Rafael, Calif. (1537 Fourth Street..."

**rdvocab.info:placeOfPublication** "San Rafael, Calif.

(1537 Fourth Street..."

**library:placeOfPublication** <http://goo.gl/maps/FaHJ>

**a schema:Place**

**a dcterms:Location**

**schema:name** "San Rafael, Calif. (1537 Fourth Street..."

**Figure 5:** Additional metadata statements inferred from an RDF map

CONTINUED »

and exploring the relationships between them, such as whether one is based on another, or extends, generalizes, or has declared equivalences with other vocabularies. This overview of the landscape, and the excellent visualization tools provided on the site, provide significant value for implementers building related services and views, as well as to the community at large identifying vocabularies at risk. The LOV project has used its research to provide recommendations for describing vocabularies so that they can be connected at the top level and viewed in relation to the larger vocabulary environment.<sup>[25]</sup>

Bernard also brought forward an initial proposal for mappings between DC properties and the schema.org vocabulary that had been announced a few months beforehand. An impromptu breakout session reviewed the first draft of those mappings and proposed a DCMI task group to flesh out and get feedback. That group is currently actively managing a prototype set of mappings using a GitHub-based project repository.<sup>[26]</sup>

## Discussion and Conclusions

Though the efforts described here represent well over a decade's worth of evolving thinking and practice, there's still a great deal to do before the vocabulary infrastructure supporting the ever-emerging Semantic Web matures sufficiently to definitively prove its worth. In the absence of top-down agreements and development planning (such absence being a "feature" of the Semantic Web in general), much of this trajectory will, of necessity, look somewhat chaotic. But given the sheer number of new and continuing efforts to expose linked data—particularly bibliographic data—the inspiration to redouble the push for supporting infrastructure that can effectively manage this chaos can't be denied.

For an example, during an update session on the Library of Congress's Bibliographic Transition Framework Initiative, Eric Miller of Zepheria<sup>[27]</sup> noted that there are now a number of projects that publish linked bibliographic data. He also noted that each of these is developing its own approach to the modeling and vocabulary selection in their data—a common practice in other early attempts to apply linked data. Recognizing that an important design feature of RDF is that metadata vocabularies are easy to define, are (optimally) self-describing to enhance interoperability, and can be used recombinantly (drawing from a variety of vocabularies in a single resource description), a relatively clear upgrade path to improvement of that data can be seen as part of the benefit of the infrastructure in the process of development.

The wide ranging conversations at the DCMI special session in The Hague remind us that interoperability and the efficiencies of common approaches require guiding principles and best practices around decisions for reuse, extension of existing vocabularies, as well as development of new vocabularies. Without cooperative efforts to develop those supportive pieces, good decisions are difficult to make, much less implement.

The role and functionality of metadata registries in the linked data infrastructure remain in flux. The requirements for macroscopic and microscopic approaches jostle for development priority, although support for vocabulary mapping functions allows a "have your cake and eat it too" balance to be maintained by ensuring that the output from both approaches is interoperable. Maps available from open registries extend the LOD environment by bringing what would otherwise be exclusively "local" vocabularies and mappings into the open domain.

It may well be that the growing interest in mapping and alignment, rather than the earlier misplaced concern around vocabulary proliferation, will fuel an important new push towards principled vocabulary practices. It's almost impossible to imagine useful Semantic mapping without well-defined, sustainable vocabularies—with that, we have the potential to move forward without impediment, leaving no parts of the community behind.

I FE I doi: 10.3789/isqv24n2-3.2012.02



**GORDON DUNSIRE** ([gordon@gordondunsire.com](mailto:gordon@gordondunsire.com)) is a freelance consultant.



**COREY HARPER** ([corey.harper@nyu.edu](mailto:corey.harper@nyu.edu)) is Metadata Services Librarian at New York University.



**DIANE HILLMANN** ([metadata.maven@gmail.com](mailto:metadata.maven@gmail.com)) is a Partner with Metadata Management Associates.



**JON PHIPPS** ([jonhipps@gmail.com](mailto:jonhipps@gmail.com)) is a Partner with Metadata Management Associates.

```

switch($row->team)
{
    case 'team':
        $query = "SELECT tea
$content = mysql_QUE
while($row = mysql_f
        {
            echo $row->team
        }
    break;

    case 'news':
        $query = "SELECT new
$content = mysql_QUE
while($row = my
        {
            echo '<ul id="news">
                <li id="title"
                <li id="conte
                <li id="contr
            </ul>';
        }
    echo '<a href="i
}
break;

case 'archive':
    query = "SELECT news
$content = mysql_QUE
while($row = my
        {
            '<ul id="news">
                <li id="title"
                <li id="conte
                <li id="contr
            </ul>';
        }
}

```

## REFERENCES

1. Isaac, Antoine, William Waites, Jeff Young, and Marcia Zeng. *Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets*. W3C Incubator Group Report, 25 October 2011. [www.w3.org/2005/Incubator/lld/XGR-ll-d-vocabdataset/](http://www.w3.org/2005/Incubator/lld/XGR-ll-d-vocabdataset/)
2. Describing Archives: a Content Standard (DACS) [www.archivists.org/governance/standards/dacs.asp](http://www.archivists.org/governance/standards/dacs.asp)
3. TEI: Text Encoding Initiative [www.tei-c.org/index.xml](http://www.tei-c.org/index.xml)
4. VRA: Visual Resources Association [www.vraweb.org/](http://www.vraweb.org/)
5. Heery, Rachel, and Manula Patel. *Application Profiles: Mixing and Matching Metadata Schemas*. Ariadne, issue 25, September 24, 2000. [www.ariadne.ac.uk/issue25/app-profiles](http://www.ariadne.ac.uk/issue25/app-profiles)
6. DESIRE Metadata Registry [desire.ukoln.ac.uk/registry/](http://desire.ukoln.ac.uk/registry/)
7. CORES Registry [www.cores-eu.net/registry/](http://www.cores-eu.net/registry/)
8. Dublin Core Metadata Registry [dcmi.kc.tsukuba.ac.jp/dcregistry/](http://dcmi.kc.tsukuba.ac.jp/dcregistry/)
9. Heery, Rachel, and Harry Wagner. A Metadata Registry for the Semantic Web. *D-Lib Magazine*, May 2002, 8(5). [www.dlib.org/dlib/may02/wagner/05wagner.html](http://www.dlib.org/dlib/may02/wagner/05wagner.html)
10. Open Metadata Registry [metadataregistry.org/](http://metadataregistry.org/)
11. Nagamori, Mitsharu, Masahide Kanzaki, Naohisa Torigoshi, and Shigeo Sugimoto. *Meta-Bridge: A Development of Metadata Information Infrastructure in Japan*. Proceedings DC-2011, The Hague, Netherlands, September 21-23, 2011, pp. 63-68. [dcpapers.dublincore.org/index.php/pubs/article/view/3632](http://dcpapers.dublincore.org/index.php/pubs/article/view/3632)
12. JISC Information Environment Metadata Schema Registry [www.ukoln.ac.uk/projects/iemsr/](http://www.ukoln.ac.uk/projects/iemsr/)
13. de hÓra, Bill. *Vocabulary Design and Integration*. Blog post of April 08, 2007. [www.dehora.net/journal/2007/04/data\\_integration.html](http://www.dehora.net/journal/2007/04/data_integration.html)
14. The RDA (Resource Description and Access) Vocabularies [rdvocab.info/](http://rdvocab.info/)
15. Willer, Mirna, Gordon Dunsire, and Boris Bosančić. *ISBD and the Semantic Web*. *JLIS.it: Italian journal of library and information science*, December 2010, 1(2): 213-236. [dx.doi.org/10.4403/jlis.it-4536](http://dx.doi.org/10.4403/jlis.it-4536)
16. *Unconstrained namespaces*. In: IFLA Namespaces Technical Group, IFLA Classification and Indexing Newsletter, no. 45 (June 2012). [www.ifla.org/files/classification-and-indexing/newsletters/IFLA-CI-newsletter-45-June2012.pdf](http://www.ifla.org/files/classification-and-indexing/newsletters/IFLA-CI-newsletter-45-June2012.pdf)
17. *British Library Data Model* [www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf](http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf)
18. COMET (Cambridge Open METadata) Code [data.lib.cam.ac.uk/code.php](http://data.lib.cam.ac.uk/code.php)
19. MARC 21 in RDF Vocabularies [marc21rdf.info/](http://marc21rdf.info/)
20. Bergman, Michael. *Bridging the Gaps: Adaptive Approaches to Data Interoperability*. Keynote Presentation at DC-2010, Pittsburgh, PA. [www.slideshare.net/mkbergman/dcmi-20101022](http://www.slideshare.net/mkbergman/dcmi-20101022)
21. Dunsire, Gordon, Diane Ileana Hillmann, Jon Phipps, and Karen Coyle. *A Reconsideration of Mapping in a Semantic World*. Proceedings DC-2011, The Hague, Netherlands, September 21-23, 2011, pp. 26-36. [dcpapers.dublincore.org/index.php/pubs/article/view/3622](http://dcpapers.dublincore.org/index.php/pubs/article/view/3622)
22. DC-2011 Vocabulary Special Session/Meeting Report [wiki.dublincore.org/index.php/DC-2011\\_Vocabulary\\_Special\\_Session/Meeting\\_Report](http://wiki.dublincore.org/index.php/DC-2011_Vocabulary_Special_Session/Meeting_Report)
23. DCMi Vocabulary Management Community [wiki.dublincore.org/index.php/DCMI\\_Vocabulary\\_Management\\_Community](http://wiki.dublincore.org/index.php/DCMI_Vocabulary_Management_Community)
24. *Linked Open Vocabularies* [lov.okfn.org/dataset/lov/](http://lov.okfn.org/dataset/lov/)
25. Vandenbussche, Pierre-Yves, and Bernard Vatant. *Metadata recommendations for linked open data vocabularies*. Version 1.0, 2011-12-28. [lov.okfn.org/dataset/lov/Recommendations\\_Vocabulary\\_Design.pdf](http://lov.okfn.org/dataset/lov/Recommendations_Vocabulary_Design.pdf)
26. DC - Schema.org Mappings [dcmi.github.com/schema.org/mappings.html](http://dcmi.github.com/schema.org/mappings.html)
27. Zepheira [zepheira.com/](http://zepheira.com/)

Jane  
Stevenson

# LINKING LIVES:

## Creating an End-User Interface Using Linked Data

JANE STEVENSON

The Archives Hub is a JISC funded service that brings together descriptions of archives held across the UK. One of the most important strengths of the Hub is the ability for researchers to make connections. They can search for people, organizations, places, or subjects across 25,000 collection descriptions and hundreds of thousands of series and item level entries. They can search serendipitously, as the index links within the Hub facilitate a lateral search that can take a user across the wealth of content so that they can discover new knowledge for their research.

In March 2010 the JISC put out a call for proposals to “expose digital content for education and research,” looking for projects that would enable structured data to be made available on the Web, in particular linked data. We secured funding for a proposal to create linked data for the Archives Hub, and the Linked Open Copac and Archives Hub (LOCAH) project was the result of this. Running over one year, it aimed to output linked data, provide views on the data, and offer a SPARQL endpoint for querying the data—as well as documenting the process through the blog. We provided a stylesheet for the transformation of Archives Hub Encoded Archival Descriptions (EAD) into RDF XML, which is available from the linked data site that we created: LOCAH Linked Archives Hub.

It seemed to us that the next logical step in the linked data journey was to create some kind of proof of concept. While the premise behind linked data is that you open up your data for others to consume and thereby provide the potential for innovative ways to combine different datasets, we felt that we needed a pro-active approach, developing our own front end—something to demonstrate the potential benefits of linked data for end users. We wanted to build on the initial investment in the LOCAH project and put linked data to the test in a real life scenario. Our proposition was that this could potentially connect archives more effectively to the wider information landscape, bringing them together with other sources to benefit researchers. It is important to state that for these reasons we wanted to have an interface

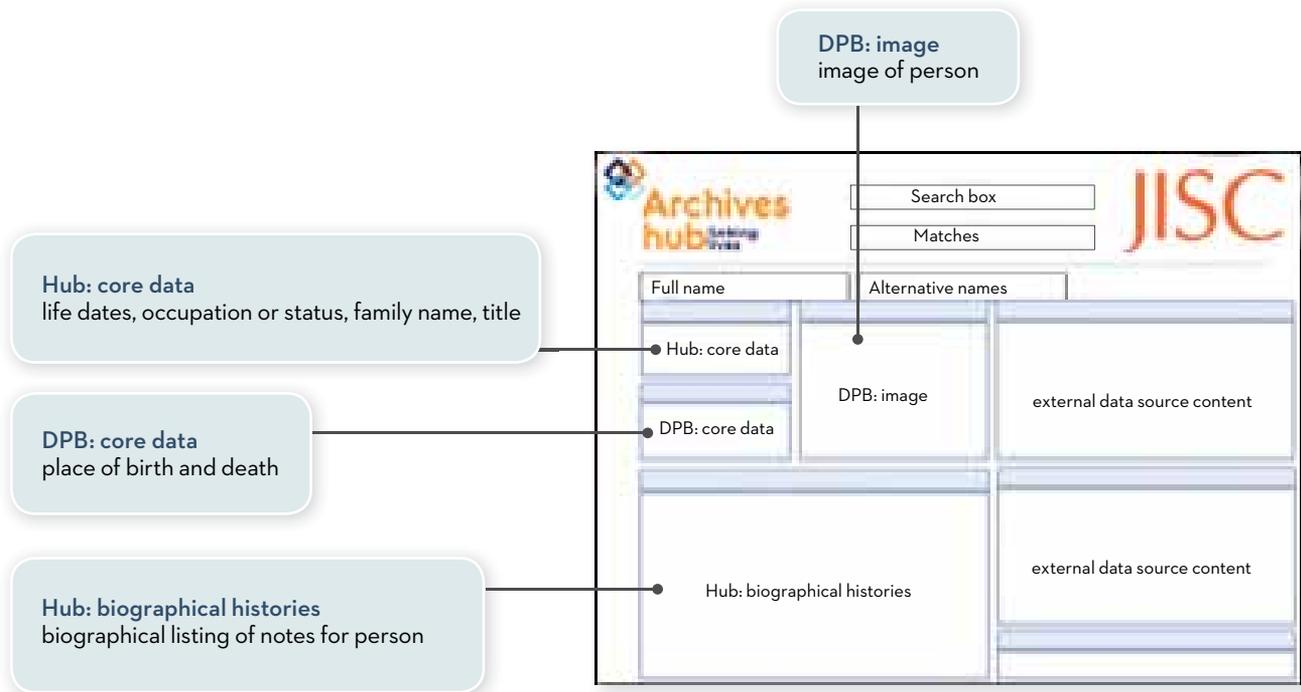


Figure 1: Wireframe for the Linking Lives interface

based entirely on linked data (that is, data provided in RDF and linked to other data sources) rather than a hybrid approach, which could include non-linked data sources.

### Why Linking Lives?

We discussed a number of ideas around which we could create an interface. The obvious options were to base it around subjects, events, or names. We decided on a biographical approach because it would clearly be of value to researchers, we felt it would be relatively easy to scope, and we had already done some matching of names within the Archives Hub to names in external datasets. Our linked data output includes statements using the <sameAs> property where we specify within our linked data that “x person in the Hub data is the same as y person in VIAF” (the Virtual International Authority File).

Linking Lives is therefore about focusing on individuals as a way into both archival collections and other relevant data sources. The Archives Hub data is rich in information about people, organizations, and events and we wanted to highlight this, as well as putting the data within the context of a range of data sources in order to provide a biographical perspective—in contrast to the more traditional interface for archives that focuses on the collection description. Researchers do not usually have an archive collection in mind when they start their research and they may not be familiar with primary sources. A biographical resource

is a familiar starting point that can lead them to relevant collections and help them to make connections between people and events.

### Interface Design

We decided to create a simple interface where one page would represent one person. We have had a number of ideas about ways to present the data and we have tried out some visualizations. But we wanted something sustainable and extensible, where we could pull in a variety of external data types—text, images, and links. Our interface uses the content boxes that are a familiar feature on many websites, and using these enables us to present different data sources as discrete parts of the interface, which is important if we want to be able to clearly identify the source of the data. (See Figure 1.)

The name appears at the top of the main display and below this a box contains key information that comes from the archive descriptions: life dates, occupation or status, family name, and title. We decided to add place of birth and death as additional core information, provided by DBPedia (see below). We placed the image in the center, as we felt this would make the interface more visually engaging. We intend to have a tab to list alternative names, which are provided by various sources, including VIAF.

We put a large box on the left-hand side to contain the all-important biographical notes for each individual that are typically created by archivists when they catalog the

CONTINUED »



Working with aggregated data from so many sources, created over a long period of time, and often migrated between different systems is a challenge. The data is inevitably inconsistent and there are errors that interfere with the data processing.

material. Beyond these key boxes, we decided that we would explore different options and experiment with the data that we could bring into the interface. This meant we did not have to decide on the final content and, indeed, it means that we can continue to add content over time, beyond the end of the project.

One of our ideas is to add an element of personalization, by enabling end users to pick and choose boxes and move them around. This remains an option, but may not be doable within the timescale of the project.

### The Challenges of the Source Data

Working with aggregated data from so many sources, created over a long period of time, and often migrated between different systems is a challenge. The data is inevitably inconsistent and there are errors that interfere with the data processing.

There are, broadly speaking, two alternative approaches to working with problematic data:

- 1 You can find ways round inconsistencies through the transformation process itself.
- 2 You can address the problems at the source.

We have written about some of the issues with the data that we have faced on our blog; the biggest issue has been with the identifiers for the archives themselves. The full identifier for the archive comprises the ISO code for the country, the UK Archon code for the repository, and the local reference for the archive—for example:

GB 983 UWA	GB Country Code	983 Repository Code	UWA Local Reference
------------	--------------------	------------------------	------------------------

On the Hub, the primary role of this reference is to be a visual indicator displayed to end users, so a level of inconsistency in the make-up of the reference within the XML document might not be a problem as long as we display it correctly; the only part the end user really needs to see is the local reference. But there is a lack of consistency in the structure of these identifiers and how the country code, repository code, and local reference are marked up in the XML. Sometimes the country code and repository code are not included and we have to work around this, but it is far harder to work with such a level of inconsistency in linked data because we want to create unique and persistent URIs out of the content.

We made the decision to go back to the Archives Hub data and construct a level of consistency, addressing any problems with duplicates and very long local references, which do not create very practical URIs. This work will be of benefit beyond the linked data project, but it is time consuming and has delayed the progress of our project somewhat.

This is only one of a number of areas where the potential for working with linked data is hampered by inconsistencies. For example, if we had standardized “extent” entries for the size of the archive, we could envisage a visualization that would show where the biggest concentrations of archives on any particular topic or person are. But these entries are highly variable because in the UK there is no recognized standard for this content, so you can have anything from “10 boxes” to “5 linear meters” to “photographs and drawings in 3 outside boxes.”

## Working with External Datasets

When working with data that comes from external sources you have no control over the data. You may have problems if it is inconsistent or if it changes. This is one of the major issues with linked data. By building an end user interface that will become part of the Archives Hub service, we should be able to get a very practical perspective on what this might mean over time.

The persistence of URIs has often been cited as an issue with linked data and although it is certainly not a problem unique to the linked data approach, it does become particularly problematic when the aim is to present a coherent and consistent information source that relies upon external URIs. So far we have not had any problems, as the URIs have been maintained, but we believe that this is an issue that needs to be monitored and assessed over time.

We have had variable success with linking to different datasets and pulling in data. To do this you need relevant content and you need the right “hooks” to pull it into the interface. We found that a number of data sources do not provide all of their data as linked data. Simply looking at the web interface can be misleading; you have to dig into the RDF and see what is there. For example, VIAF provides a list of selected titles for authors, but this information is not included within the linked data. In addition, some data sources do not provide a SPARQL interface, which is what is typically used to query data. So far we have struggled to find linked data that includes connections between people; for example, a simple statement that “x person knows y person.” Our hope was to include these types of relationships as we wanted to build up a resource that would show connections between people.

We created our own Wiki in order to list different datasets and provide summary notes about them. Datasets we have looked at include DBPedia, OpenLibrary, VIAF, Freebase, BBC Programmes, and Linked Open British National Biography (BNB). It is unlikely that we will be able to add data from all of the datasets we assess within this project, even if they all have relevant and useful data, because of time constraints. But we can continue to use the Wiki to monitor potential data sources and add them at a later date. We may also make the Wiki public in order to share our experiences and findings.

We agreed from the outset that we wanted to bring in data from Wikipedia (DBPedia being the linked data version of Wikipedia). But, as with many other external datasets, we have hit one significant problem: not all records on Wikipedia have the same information. So, for example, we have provided for space for an image of the individual, but we will not always have that image available. We are



We have had variable success with linking to different datasets and pulling in data. To do this you need relevant content and you need the right “hooks” to pull it into the interface. We found that a number of data sources do not provide all of their data as linked data.

---

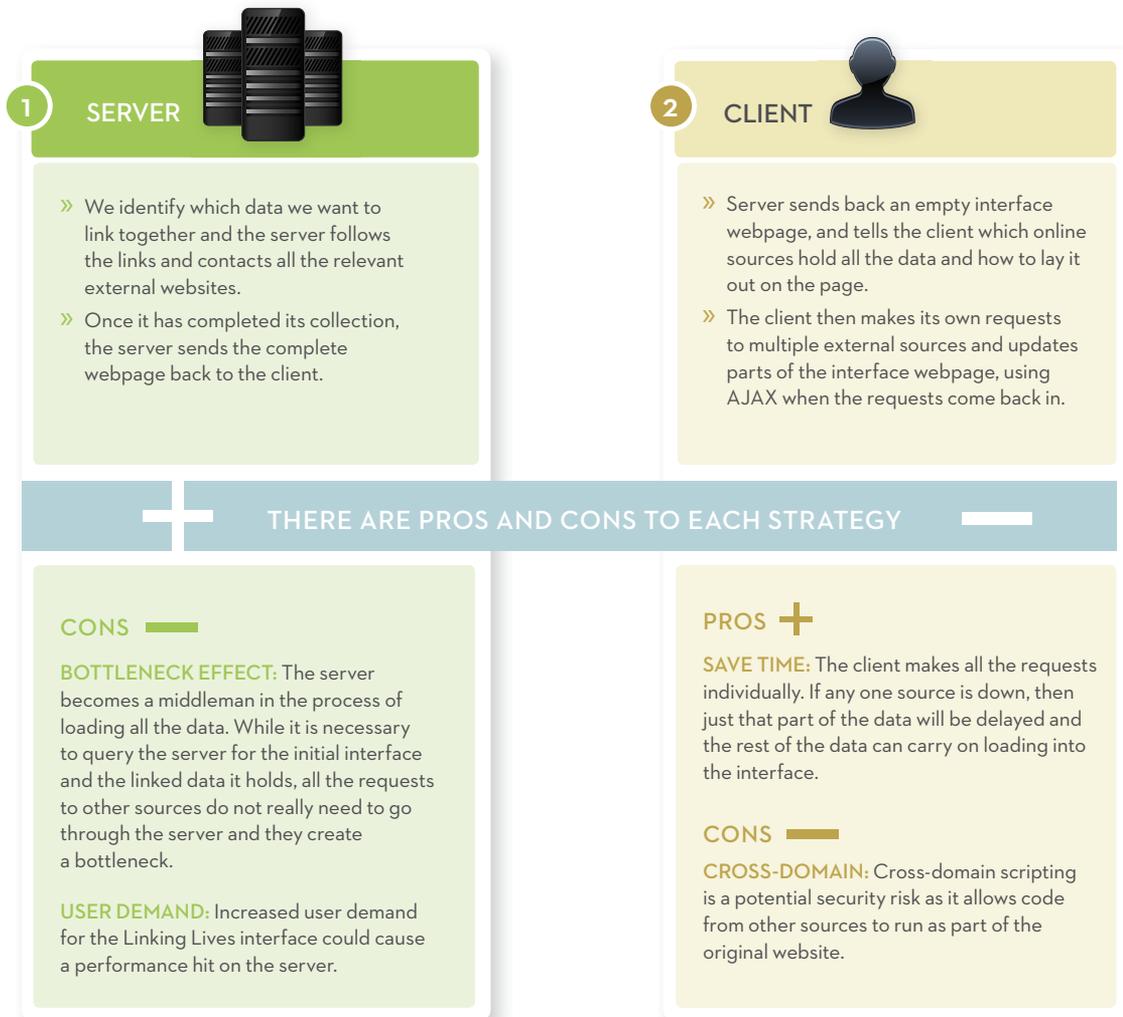
considering options for ways to address this issue, and we may take the same approach as the BBC, which includes Wikipedia content in its webpages (See for example: <http://www.bbc.co.uk/nature/life/Felidae>). The BBC makes clear where the content is from and invites readers to edit the Wikipedia article.

## Understanding the Interface

With our interface, we want to show that archives can benefit from being presented not in isolation, but as a part of a fuller picture—alongside different data sources—to create a rich biographical resource. People do not always find dedicated archives sites easy to use. The hierarchical nature of archives and the nature of collections (which can be anything from one item to a vast collection of items in different media) can make them difficult to represent online. Combining them with other sources and presenting them in a different way may facilitate interpretation, but it is essential to evaluate this hypothesis, to find out how researchers react to what they are presented with, and whether they believe it is useful for their work.

CONTINUED »

## TWO STRATEGIES THAT CAN BE EMPLOYED FOR COLLECTING DATA TOGETHER WITHIN LINKING LIVES:



We have a group of students and researchers from The University of Manchester taking part in an evaluation of the Linking Lives interface. We wanted to ascertain their thoughts about the more traditional archival interface and get a sense of their understanding of archives, so initially we asked them to visit the Archives Hub and give us their thoughts in response to a number of questions. Our intention now is to run a focus group with these participants where we introduce them to the new interface. We intend to incorporate their feedback into a modified design.

Aside from bringing together different data sources, one of the features of the new interface is that it brings together a number of biographical histories for any one person, if that

person has created more than one archive. We are particularly interested to find out how researchers react to this: whether they find it useful and whether the inevitable repetition of information is seen as a distraction.

### A Technical Perspective

The Linking Lives interface is a web application loaded onto a user's web browser (the client) from our server.

As such, there were two obvious strategies we could employ for collecting the data together within the application:

- 1 Let the server do the data collection.** In this scenario, we identify which data we want to link together and the server

follows the links and contacts all the relevant external websites. Once it has completed its collection, the server sends the complete webpage back to the client.

- ② **Let the client web browser do the data collection.** In this scenario, the server sends back an empty interface webpage, and tells the client which online sources hold all the data and how to lay it out on the page. The client then makes its own requests to multiple external sources and updates parts of the interface webpage, using AJAX when the requests come back in.

There are pros and cons to each strategy. In the first scenario, the server becomes a middleman in the process of loading all the data. While it is necessary to query the server for the initial interface and the linked data it holds, all the requests to other sources do not really need to go through the server and they create a bottleneck. One notable effect of such a bottleneck would be that if an external data source was down, or performing slowly, the loading of the entire interface would be delayed while the server waited for the response. It is also conceivable that increased user demand for the Linking Lives interface could cause a performance hit on the server. Another consequence of using the server in this fashion is that it must decide on all the queries it is going to make of remote sources and run them before the user sees anything.

In the second scenario, the client makes all the requests individually. If any one source is down, then just that part of the data will be delayed and the rest of the data can carry on loading into the interface. Additional queries can be made on the fly; information from one source can be used to generate a query on another source, or even to amend and update a query that has already run. All of this can be going on while the user has something to look at on the screen.

The problems with the second scenario come in the form of increased complexity in the interface logic and the cross-domain problem. Cross-domain scripting is a potential security risk as it allows code from other sources to run as part of the original website. Sites may not have the capability to accept requests like this, or they may block webpages that try to load further content from other webpages. This problem could potentially make this solution untenable, which may be a significant problem for an open linked data approach.

There are a number of workarounds to the cross-domain problem. The W3C have a recommended solution involving remote websites supplying an extra piece of header information that confirms that data from their page can be loaded into other pages as long as it is properly requested. As long as this feature—known as Cross-Origin Resource Sharing (CORS)—is enabled on remote servers, the second scenario is possible.

We took the decision to implement this second option, despite the extra work involved, as it provided for a more flexible and effective solution and makes the design more open ended.

## Problems of Identity

One of the biggest challenges around our linked data work has been identifying individuals—a particular focus for us because Linking Lives is based upon people. The URIs used to identify persons in the Linked Archives Hub dataset have their origins in the names of persons occurring in the Archives Hub EAD XML documents.



One of the biggest challenges around our linked data work has been identifying individuals; a particular focus for us because Linking Lives is based upon people. The URIs used to identify persons in the Linked Archives Hub dataset have their origins in the names of persons occurring in the Archives Hub EAD XML documents.

CONTINUED »



As different forms of the name can legitimately be used to refer to the same person, our current transformation process means that we end up with multiple URIs for one individual.

Within those documents, person names occur in two contexts:

### 1 *Personal names as index terms*

The first context is that of personal names added to the description by the cataloger as index terms, on the basis that they may be useful for the purposes of retrieval/search/browse.

An index term for one individual may occur several times within the Archives Hub data. For example, *Webb, Martha Beatrice, 1858-1943, social reformer*, occurs in three different EAD XML documents. This name is taken from the National Register of Archives held in the UK (the NRA), so this is cited as the source of the index term.

For this term the URI is:

```
http://data.archiveshub.ac.uk/id/person/nra/webbmarthabeatrice1858-1943socialreformer
```

Use of the rules may lead to different descriptors being used for the person, so we have the URIs:

```
http://data.archiveshub.ac.uk/id/person/nca/rules/webbmarthabeatrice1858-1943neepottersocialreformerandhistorian
```

```
http://data.archiveshub.ac.uk/id/person/nca/rules/webbmarthabeatrice1858-1943socialreformer
```

These use the UK National Council on Archives (NCA) Rules. As different forms of the name can legitimately be used to refer to the same person, our current transformation process means that we end up with multiple URIs for one individual.

In addition to this, use of the name within the URI does not avoid any issues of ambiguity. It is very unlikely with a name like *Martha Beatrice Webb*, but it is very possible with many names within archive descriptions, as they do not always include life dates and so you may have something like *Mary Jones, b 1901* and *M Jones, 1901-1980* in two different archive descriptions, both adhering to the same rules for name construction and referring to the same person. You may also have *John Smith, b 1945, engineer* in two different descriptions, which would create the same URI, but it may not be the same person.

A further problem is that names may change when death dates are added. This means the subsequent re-transformation of the data will generate a different URI from that generated by the previous process using the initial form of the name.

### 2 *Personal names as creators*

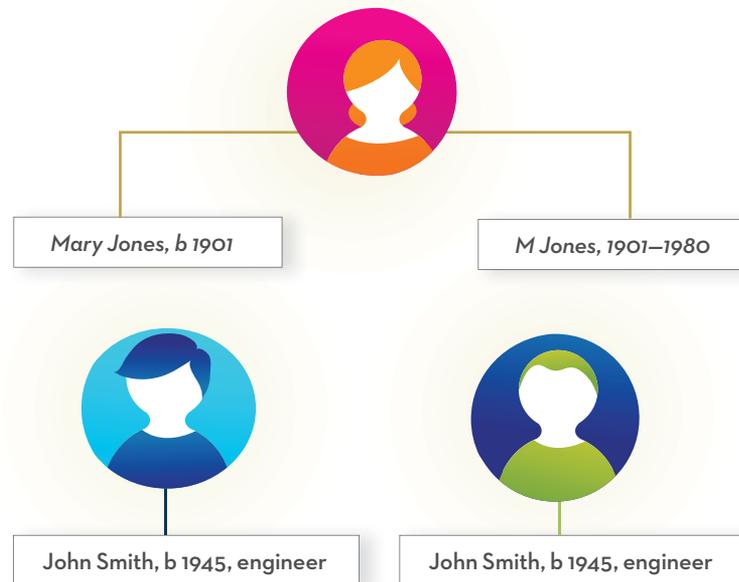
Personal names are also found within an EAD entry for the name of creator (or originator) of the archive—the agent(s) responsible for the creation or bringing together of the resources described. In the Hub EAD data, the names are not marked up to distinguish the name of a person from that of an organization. Furthermore, this entry is a free text entry, and usually the commonly used form of the name is given, so it is not easy to map this to the index entry.

An example of a URI generated for this data is:

```
http://data.archiveshub.ac.uk/id/agent/gb97/webbmarthabeatrice1858-1943wifeof1stbaronpassfieldsocialreformerandhistorian
```

In addition to this, use of the name within the URI does not avoid any issues of ambiguity.

So you may have something like *Mary Jones, b 1901* and *M Jones, 1901-1980* in two different archive descriptions, both adhering to the same rules for name construction and referring to the same person. You may also have *John Smith, b 1945, engineer* in two different descriptions, which would create the same URI, but it may not be the same person.



We include the repository reference (gb97), so that the name is effectively the person as represented within that repository.

We used a number of processes to identify candidate matches within the Hub dataset between “agents” (generated from the creator/origination context) and “persons” (generated from the index terms context). A degree of manual checking was then used to assess the accuracy of these candidate matches before creating “sameAs” relationships to indicate that the URIs refer to the same person.

One of the problems with this approach is that an application consuming the data still has to be prepared to work with these multiple URI aliases and, particularly with SPARQL, this can be quite cumbersome. To find all the data we hold about the person denoted with URI X, an application has to search for patterns involving not just that known URI X, but also any URI Y, where URI Y is “sameAs” URI X.

One approach to the repeatability problem would be to see the transformation stage as only the first part of a larger process, to keep track of the URIs generated over time, and build in a stage of processing to reconcile the URI generated this week from *Scott, James, 1950-2012, Sir, biologist* from the URI generated from *Scott, James, 1950-, scientist* in the previous version of the document six months ago. This perhaps then becomes simply a special case of dealing with multiple URIs for a single entity.

To avoid multiple URIs for one individual, it may well be that rather than publishing a set of “sameAs” triples, we should take a step further and consider consolidating our data to use a single URI for the person. But which version do we distill our multiple URIs down to? Or do we create a new URI for the individual? Should we instead think about creating a mapping to some sort of code and use that to construct a distinct URI? Or maybe more radically, but potentially more practical, would be to use existing external URIs in our data, such as the URIs from the VIAF name authority. However, it is unlikely that any resource would be able to provide URIs for all of the names in the Archives Hub dataset. In addition, there would be issues with control and persistence, as well as “dereferencing” the URI to provide information about the entity. But there would certainly be advantages to the principle of using the same URIs for the same entity across different datasets.

The issues surrounding identification of persons are many and complex. Our Linking Lives project has helped us to understand the practical implications of using our linked data, but we are not yet in a position to say that we have found a sustainable and reliable way to identify individuals. This is not ideal when you are trying to make something work in a practical, cost-effective way.

CONTINUED »

## Conclusions

Part of the motivation behind Linking Lives is to assess whether linked data really does provide an alternative way forward. We believe that we are creating a useful and valuable resource, and we are successfully connecting to external datasets using Linked Data principles. Linking Lives enables us to give archives a different context, putting them into a broader knowledge domain, and we will be able to evaluate the response to this approach from researchers. Our hope is that it provides a useful case study for others who are undertaking similar projects.

We have continued to find linked data work challenging, partly due to the fact that it is a new and developing area with few templates or tools to utilize, partly due to the challenges of working with various external data sources, and partly because of issues within our own data. We needed to take a lightweight approach to project management and to adopt an iterative technical development methodology because it was difficult to set clear objectives.

With limited time and resources for what turned out to be a more complex project than we had initially envisaged, we necessarily had to prioritize. One decision we made was to focus on the interface, rather than the search and navigation elements of the service. If the interface proves to be useful to end users, we will continue to develop the search capability and look to integrate it more fully with the main Archives Hub service.

I would say that the biggest single factor in terms of additional work has been cleaning up our own data. The inconsistencies within data created by so many institutions over such a long period are compounded by the complex nature of hierarchical EAD finding aids. This work requires a level of expertise in archival description as well as specialist skills in linked data.

We did not have time to look in detail at as many external datasets as we would have liked, but more than this, the linked data space is constantly changing, so new data is created all the time and improvements are made to existing data. This makes it quite a moveable feast, and you have to make decisions about whether to go back to updated datasets and re-examine them, or stick with what you have. This may be a challenge in terms of maintaining the interface. We may find that the need to monitor the linked data space takes up significant time. We will continue to maintain our linked data interface and seek to add some more external data sources, and then we will monitor the result, see how much it is used, and how much effort we have to invest in ensuring it is current and all links are operable.

My feeling is that it needs to be easier to locate and probe data sources to ascertain the classes of things being described and the properties used to describe them—and it needs to be easier to link to these external sources. The CKAN Data Hub is one attempt to bring data together, but it is not comprehensive and not entirely easy to navigate. However, it must be recognized that working with open data in this way is not going to be easy, and a degree of investigation may be necessary to establish exactly what is being provided and how uniform it is. Simply connecting and cross-searching just two datasets using more traditional means can often prove to be challenging; with Linked Data the idea is to be able to access and connect numerous open datasets in RDF.

A tall, slightly irregular stack of many papers or documents, some with visible edges and colors. To the left of the stack is a circular icon containing a magnifying glass, symbolizing search or investigation.

I would say that the biggest single factor in terms of additional work has been cleaning up our own data. The inconsistencies within data created by so many institutions over such a long period are compounded by the complex nature of hierarchical EAD finding aids.

With big players like the Library of Congress committing more fully to linked data with the Bibliographic Framework project, a certain level of optimism in the promise of linked data is clearly still in evidence, and the community is continuing to expand and evolve. There also seems to be significant and increasing interest from the LOD-LAM community (Linked Open Data for Libraries, Archives, and Museums). However, there are indications that linked data is still evolving too slowly to attract the level of investment necessary to make it a viable business enterprise and attract significant investment. (See the blog post by Tim Hodson of Talis.) Does the altruistic goal of opening up data to advance knowledge and benefit research provide a strong enough impetus to drive the linked data ideal?

IP | doi: 10.3789/isqv24n2-3.2012.03

**JANE STEVENSON** ([jane.stevenson@manchester.ac.uk](mailto:jane.stevenson@manchester.ac.uk)) is Archivist and Archives Hub Manager at Mimas, based at The University of Manchester in the UK.

*This article was written with help from Adrian Stevenson and Lee Baylis (Mimas) and Pete Johnston (University of Cambridge)*



Simply connecting and cross-searching just two datasets using more traditional means can often prove to be challenging; with Linked Data the idea is to be able to access and connect numerous open datasets in RDF.

#### Archives Hub

[archiveshub.ac.uk/](http://archiveshub.ac.uk/)

#### ARCHON Directory

[www.nationalarchives.gov.uk/archon/](http://www.nationalarchives.gov.uk/archon/)

#### BBC Programmes

[www.bbc.co.uk/programmes](http://www.bbc.co.uk/programmes)

#### British National Biography (BNB) Linked Open Data

[thedatahub.org/dataset/bluk-bnb](http://thedatahub.org/dataset/bluk-bnb)

#### CKAN Data Hub

[thedatahub.org](http://thedatahub.org)

#### Cross-Origin Resource Sharing (CORS)

[www.w3.org/TR/cors/](http://www.w3.org/TR/cors/)

#### Datalinks Wiki (Archives Hub Linked Open Data)

[datalinks.wikia.com/wiki/Archives\\_Hub\\_Linked\\_Data](http://datalinks.wikia.com/wiki/Archives_Hub_Linked_Data)

#### DBpedia

[dbpedia.org/about](http://dbpedia.org/about)

#### Freebase

[www.freebase.com/](http://www.freebase.com/)

#### JISC

[www.jisc.ac.uk](http://www.jisc.ac.uk)

#### Library of Congress Bibliographic Framework Project

[www.loc.gov/marc/transition/news/framework-103111.html](http://www.loc.gov/marc/transition/news/framework-103111.html)

#### Linked Open Data for Libraries, Archives, and Museums

[lodlam.net](http://lodlam.net)

#### LOCAH Linked Archives Hub

[data.archiveshub.ac.uk/](http://data.archiveshub.ac.uk/)

#### OpenLibrary

[openlibrary.org](http://openlibrary.org)

#### SPARQL Query Language for RDF

[www.w3.org/TR/rdf-sparql-query/](http://www.w3.org/TR/rdf-sparql-query/)

#### Too early, too slowly. [timhodson.com](http://timhodson.com) [blog]. July 5, 2012

[timhodson.com/2012/07/too-early-too-slowly/](http://timhodson.com/2012/07/too-early-too-slowly/)

#### Virtual International Authority File (VIAF)

[viaf.org/](http://viaf.org/)



RELEVANT  
LINKS

Seth van  
HoolandRuben  
VerborghRik Van  
de Walle

# Joining the Linked Data Cloud *in a Cost-Effective Manner*

SETH VAN HOOLAND, RUBEN VERBORGH, AND RIK VAN DE WALLE

Linked Data hold the promise to derive additional value from existing data throughout different sectors, but practitioners currently lack a straightforward methodology and the tools to experiment with Linked Data. This article gives a pragmatic overview of how general purpose Interactive Data Transformation tools (IDTs) can be used to perform the two essential steps to bring data into the Linked Data cloud: data cleaning and reconciliation. These steps are explained with the help of freely available data (Cooper-Hewitt National Design Museum, New York) and tools (Google Refine), making the process repeatable and understandable for practitioners.

## Linked Data comes at a cost

Many institutions are now aware of the importance of having their data available as Linked Data. The five-star scheme proposed by Tim Berners-Lee seems a valuable tool to assess the reusability of data for current and future applications. However, some goals are more difficult to reach than others: for example, linking to other data is currently a rare practice, yet this is crucial to provide context for both human and machine data consumers. The evaluation scheme unfortunately makes abstraction of the quality of the data: while most institutions have data available in a structured format, consistency issues within individual data fields

present tremendous hurdles to create links in between datasets in an automated manner.

Streamlining and cleaning data to enhance the linking process in between heterogeneous data sources used to be a task that had to be performed by people with both high domain and technological skills. Often, many thousands of records require similar operations. This is either a tedious manual task or something that needs to be automated on a per-case basis. Luckily, the advent of Interactive Data Transformation tools (IDTs) allows for rapid and inexpensive operations on large amounts of data, even by domain experts who do not have in-depth technical skills. But exactly how much can be achieved with IDTs and how

reliable are the results? These are the questions we have been investigating in the scope of the Free Your Metadata initiative. We were able to verify that IDTs can assist with cleaning and linking of large datasets, leading to a high success percentage at a minimal cost. (For more on this, see our forthcoming JASIST article: *Evaluating the success of vocabulary reconciliation for cultural heritage collections*. Pre-print at: [freeyourmetadata.org/publications/](http://freeyourmetadata.org/publications/))

### The Interactive Data Transformation revolution

Interactive Data Transformation tools resemble the desktop spreadsheet software we are all familiar with. While spreadsheets are designed to work on individual rows and cells, IDTs operate on large datasets at once. These tools offer a homogeneous and non-expert interface through which domain experts can perform both the cleaning and reconciliation operations. Several general-purpose tools for interactive data transformation have been developed over the last years, such as Potter’s Wheel ABC and Wrangler.

Here we want to focus specifically on Google Refine (formerly Freebase Gridworks), as it has recently gained a lot of popularity and is rapidly becoming the tool of choice to efficiently process and clean large amounts of data in a browser based interface. Google Refine further allows the reconciliation of data with existing knowledge bases, creating the connection with the Linked Data vision. The DERI research group has developed an RDF extension for Google Refine, which can be downloaded for free. The RDF extension allows users to add SPARQL endpoints to the reconciliation process. DBpedia is added, for example, so that the content of a keyword type of field can be matched to terms described as SKOS concepts in DBpedia. More specialized sources such as the LC Subject Headings (LCSH) and the Art & Architecture Thesaurus (AAT)® can also be used.



#### CASE STUDY



Cooper-Hewitt National Design Museum, New York

The use of Google Refine for metadata cleaning and reconciliation will be demonstrated with the help of the metadata of the Cooper-Hewitt National Design Museum, which released its collection metadata as a downloadable file through GitHub in February 2012, using the Creative Commons Attribution Share Alike (CCSA) license. The file contains basic metadata (29 fields) for 137,571 objects.

The cleaning operations are performed on the entire dataset but for reasons of simplicity and performance the reconciliation process specifically focuses on the metadata record for *Design for a candelabrum* by Michelangelo, as it is one of the iconic objects from the collection. The drawing is available through the Google Art project.



Cooper-Hewitt National Design Museum  
*Design for a candelabrum*  
by Michelangelo

CONTINUED »



Figure 1: Clustering with Google Refine to allow the detection of terms with inconsistencies

# 1

## Data cleaning

Once the data are imported into Google Refine, a diverse set of filters and facets can be applied on the individual fields. All manipulations are performed through a clear and straightforward interface, allowing domain experts without any technical background to experiment with data normalization and cleaning.

The following operations illustrate some of the most recursive issues with data and how Google Refine can be used to both identify, and where possible, solve them in an automated manner.

### Deduplication

After loading the data into the application, the first operation we need to perform is to detect and remove duplicates. This can easily be done by performing the *Duplicates facet* on fields such as *objectid* and *invno* (inventory number). For example, 6,215 records were identified through this facet that have a duplicate inventory number.

### Atomization

A quick glance at the *medium* field, which typically has content such as “Quill-work, silver, glass, and black-painted pine,” or the content of the *geography* field (e.g., “London England”), illustrates one of the biggest hurdles for automated data analysis and reconciliation: field overloading. These values need to be split out into individual cells through the function *Split multi-valued cells* on the basis of separation characters, which are a comma and a whitespace in the case of the *medium* field and a whitespace in the case of the *geography* field.

## Applying facets and clustering

Once the content of a field has been properly atomized, filters, facets, and clusters can be applied to give a quick and straightforward overview of classic formal data issues. By applying the custom facet *facet by blank*, one can in a matter of seconds measure the completeness of the fields; for example, 72% of the *description* fields and 93% of the *movement* fields from the Cooper-Hewitt collection are left blank.

The *text facet* is one of the most powerful features of Google Refine, as it instantly identifies both the most recurrent values and the outliers of a field. When applied on the *names* field for our collection, we see a total number of 3,785 different values composed of a small number of terms that are heavily used (e.g., “drawing” is used to describe 27% of the objects) and a long tail of object names which are only used once. After the application of a facet, Google Refine proposes to cluster facet choices together based on various similarity methods, such as nearest neighbor or key-collision. The two or more related values are presented and a merge is proposed, which can either be approved or manually overridden. Figure 1 illustrates the clustering and how it allows resolution of case inconsistencies, incoherent use of either the singular or plural form, and simple spelling mistakes. However, a manual check of the proposed clusters is necessary as attention needs to be given to near-duplicates such as *toaster* – *coaster*. The application of the nearest neighbor clustering method, which is considered as the least aggressive, typically reduces the number of variant values by 10 to 15%.

# 2

## Reconciling data with the Linked Data cloud

Once the data has been cleaned, the moment has come to give *meaning* to the field values. We as humans understand what “Italian” and “Renaissance” mean, but to machines both terms are just strings of characters. With Linked Data, meaning is created by providing *context* in the form of *links*. For example, for “Renaissance,” we mean the cultural movement in Europe during the 14th to 17th centuries, as defined by the link <http://en.wikipedia.org/wiki/Renaissance>. The process of matching text strings to concepts is called *reconciliation* in Google Refine. Reconciliation can be performed automatically on all records on a per-field basis. Among interesting columns to reconcile in the Cooper-Hewitt collection are name, culture, and period. For each column, you can specify the reconciliation source, and the type of entity contained in the column (see Figure 2).

We will illustrate reconciliation on the object “Design for a Candelabrum.” If we reconcile the *name* field with the LCSH vocabulary, the value “Drawing” becomes linked to the

LCSH concept *Drawing*. The latter is more than simply a string; it is a concept in a hierarchy, with relations to other terms. The word “Italian” can be reconciled automatically to the Freebase entry of *Italy*. If we try to reconcile “Late Renaissance” with the LCSH, Refine offers us two alternatives between which it cannot choose automatically: *Art, Late Renaissance* and *Painting, Late Renaissance*. While we need to select our choice manually, Refine does limit the number of choices we have to make.

The links that result from the reconciliation process not only help machines, they also eventually help people consume information faster and smarter. For example, if the *maker* field is reconciled to the *Michelangelo* article in Wikipedia, people have access to relevant information directly. If many items from different collections are linked this way, people can browse related works automatically. Reconciliation thereby connects each collection to the Linked Data cloud.

CONTINUED »

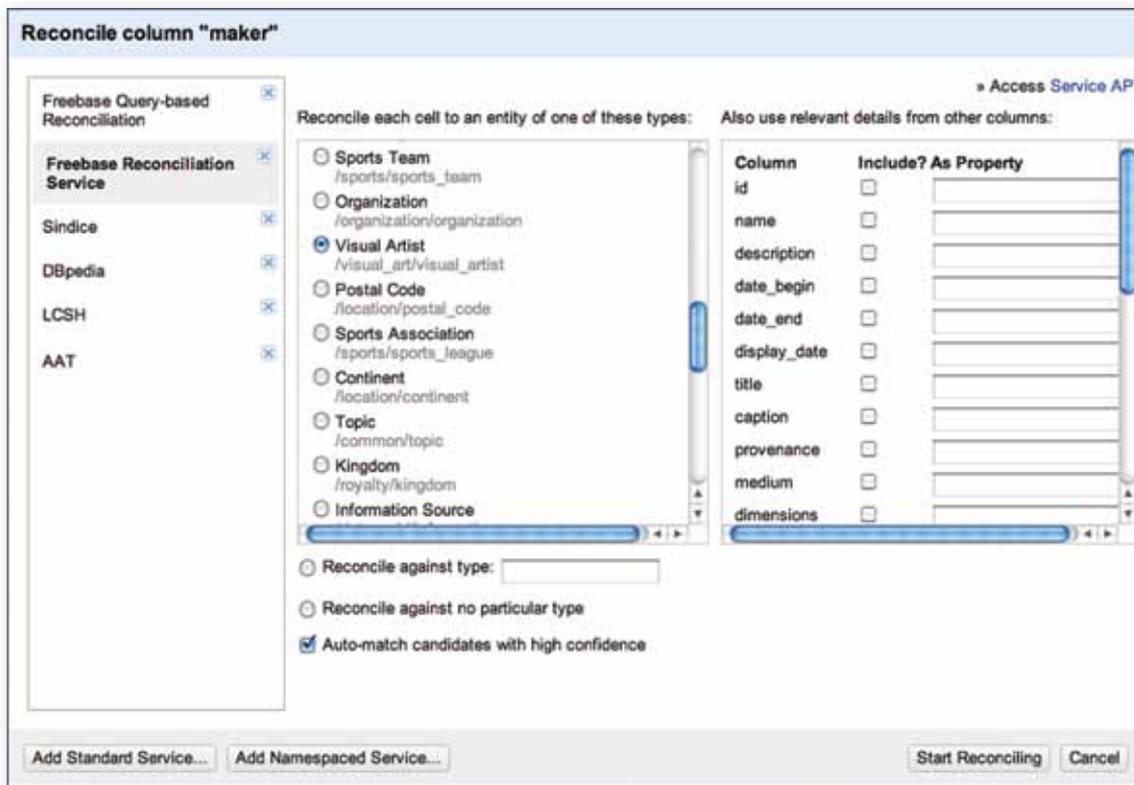


Figure 2: Reconciling columns of data with Google Refine using different sources and suggestions of the entity type to look for



## Start picking the low hanging fruit!

With the help of freely available data and tools, we demonstrated in a straightforward manner how non-technical people can bring their own data into the Linked Data cloud. The arrival of IDTs and Google Refine, in particular, has made data cleaning and reconciliation available for the masses. Concrete examples showed how recurrent data quality issues can be handled by Google Refine and how to transform strings of text into links pointing to external data sources that are already a part of the Linked Data cloud.

The quickly evolving landscape of standards and technologies certainly continues to present challenges to non-technical domain experts wishing to derive additional value out of their data on the Web. We do not wish to make light of the inherent complexities involved in the interlinking of data, but we do want to point out the low hanging fruit that is currently right in front of our noses. This small case study demonstrates that significant results can be obtained at a minimal cost through freely available tools.

## Acknowledgements

*The authors would like to thank the Cooper-Hewitt National Design Museum for making their metadata freely available and therefore allowing us to perform the case study on the basis of metadata which can be used under the CCASA license.*

*The research activities described in this paper were partly funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union. | IP | doi: 10.3789/isqv24n2-3.2012.04*

---

**SETH VAN HOOLAND** (svhoolan@ulb.ac.be) is Digital Information Chair at Université Libre de Bruxelles. **RUBEN VERBORGH** (ruben.verborgh@ugent.be) is a PhD Student at Ghent University, Interdisciplinary Institute for Broadband Technology (IBBT) Multimedia Lab. **RIK VAN DE WALLE** (rik.vandewalle@ugent.be) is Senior Full Professor at Ghent University and Research Director of the Interdisciplinary Institute for Broadband Technology (IBBT) Multimedia Lab.

### Art & Architecture Thesaurus (AAT)®

[www.getty.edu/research/tools/vocabularies/aat/index.html](http://www.getty.edu/research/tools/vocabularies/aat/index.html)

### Berners-Lee, Tim. *Linked Data. [Is your Linked Open Data 5 Star?]*

[www.w3.org/DesignIssues/LinkedData.html](http://www.w3.org/DesignIssues/LinkedData.html)

### Cooper-Hewitt National Design Museum Labs

[labs.cooperhewitt.org/](http://labs.cooperhewitt.org/)

### Cooper-Hewitt's collection metadata

[labs.cooperhewitt.org/2012/releasing-collection-github/](http://labs.cooperhewitt.org/2012/releasing-collection-github/)

### DBpedia

[dbpedia.org/](http://dbpedia.org/)

### Free Your Metadata

[freeyourmetadata.org/](http://freeyourmetadata.org/)

### Freebase

[www.freebase.com/](http://www.freebase.com/)

### Freebase - Italy entry

[www.freebase.com/view/en/italy](http://www.freebase.com/view/en/italy)

### Google Refine

[code.google.com/p/google-refine/](http://code.google.com/p/google-refine/)

### LC Subject Headings (LCSH)

[id.loc.gov/authorities/subjects.html](http://id.loc.gov/authorities/subjects.html)

### LCSH Subject Headings - Drawing entry

<http://id.loc.gov/authorities/subjects/sh85039408.html>

### Michelangelo. *Design for a candleabrum.*

<http://www.googleartproject.com/collection/cooper-hewitt-national-design-museum/artwork/design-for-a-candelabrum-michelangelo/12466224/>

### Potter's Wheel ABC

[control.cs.berkeley.edu/abc/](http://control.cs.berkeley.edu/abc/)

### RDF Extension for Google Refine

[lab.linkeddata.deri.ie/2010/grefine-rdf-extension/](http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/)

### SPARQL Endpoints

<http://www.w3.org/wiki/SparqlEndpoints>

### Wrangler

[vis.stanford.edu/papers/wrangler](http://vis.stanford.edu/papers/wrangler)



RELEVANT  
LINKS



# OCLC's Linked Data Initiative: *Using Schema.org to Make Library Data Relevant on the Web*

TED FONNS, JEFF PENKA, AND RICHARD WALLIS

In June of 2012, OCLC announced the next stage of its Linked Data strategy when it revealed that Schema.org markup had been added to WorldCat.org pages under an Open Data Commons license (ODC-BY). This technique provided a platform to present the metadata and holdings for millions of bibliographic items held by tens of thousands of libraries to the large commercial search engines for use in their search indexes and applications.



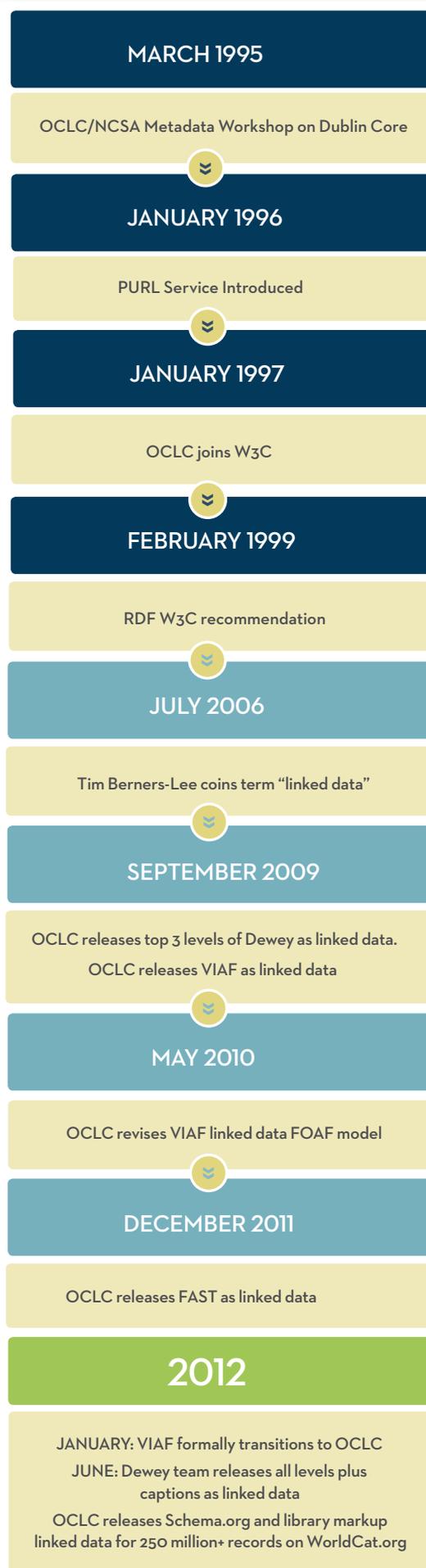
The Schema.org initiative—made up of Google, Bing, Yahoo, and Yandex—provides a core ontology for search engines and other web crawlers to directly make use of this library data. Schema.org represents a cooperative agreement between these major search engines to share a core vocabulary for markup. It helps the search engines to normalize the markup of webpages in a way that reduces ambiguity about what the pages are describing and makes the integration of the data into search engines more efficient.

OCLC observed this development in the search engine industry and realized that it could be an important tool to more effectively represent the collective collections of libraries on the Open Web. At the same time, OCLC's

internal experiments with linked data were maturing, and the opportunity to combine this new method for exposing data on the Web with the value of library linked data seemed ideal. OCLC enhanced the core bibliographic data exposed on WorldCat.org with Schema.org markup and, following good linked data practice, included Universal Resource Identifiers (URIs) for as many linkable data elements as possible.

In the complex message exchange between a web user's search and the content to be delivered (library collections in this case), Schema.org markup provides an ideal tool to mediate that complexity and more efficiently connect end users to the content they desire.

CONTINUED »



## OCLC LINKED DATA TIMELINE

CONTINUED »

### How did we get here?

OCLC's interest in providing structured data suitable for wide consumption goes back to the Dublin Core initiative in 1995 when OCLC hosted a meeting of international experts at its headquarters in Dublin, Ohio to develop a core vocabulary for the description of resources. In 1997, OCLC joined the W3C, and staff in OCLC Research became active participants in the subsequent discussions on how best to represent library data on the Web. The late 2000s saw OCLC begin to experiment with the benefits of exposing library linked data through a series of experimental releases. In 2009, OCLC released the top three levels of Dewey as linked data through Dewey.info. Also in 2009 the Virtual International Authority File (VIAF) was released as linked data. That release was improved in 2010 with a Friend of a Friend (FOAF) model release of VIAF.

The release of VIAF as linked data represented a powerful opportunity to provide durable and authoritative data about authors and titles on the Web in a way that encourages linking to library resources. In 2011, OCLC released the Faceted Subject Terms (FAST) data as linked data to provide a controlled subject vocabulary to the linked data environment. More recently, in 2012, the evolution of Dewey.info moved forward significantly with the release of all levels and captions of the Dewey controlled subject vocabulary.

Point-of-need access for web users drove OCLC's introduction of WorldCat.org in 2005. By surfacing the collective collection of the world's libraries and working with partners like Google, Bing, and Yahoo!, millions of web users now have rich library content appearing in their regular workflows. Given the variety of linked data work at OCLC and the goals of WorldCat.org, the Schema.org effort offered a great opportunity for a webscale exercise to bring it all together.

To further improve the representation of library data on the Web, OCLC is working with the Schema.org community to develop and add a set of vocabulary extensions to WorldCat data. Schema.org and library-specific extensions will provide a valuable two-way bridge between the library community and the consumer web.

### The Technical Process (The Nuts and Bolts)

Meaningful Schema.org-derived linked data was added to WorldCat.org in three phases.

- 1 The OCLC linked data team focused on data modeling necessary to connect existing experimental linked data projects (e.g., VIAF, FAST, LC Authorities, and Dewey) to the Schema.org base vocabulary and created an initial library extension to the vocabulary.



- 2 The team experimented with various data models and approaches to apply descriptive, linked data decoration to the bibliographic content on WorldCat.org. The data-intensive nature of this iterative process required technology that handles the variety and volume of data along with the iterative process of setting models, running them against the data, reviewing the results, and adapting the models. These requirements for rapid iteration were addressed by the use of the Apache Hadoop software framework, which shortened the data-loading time for hundreds of millions of records from weeks to minutes.
- 3 The final stage required the updating and displaying of linked data-decorated WorldCat.org records on the production site for use by web users, partners, and harvesters. The WorldCat.org site is optimized for high-traffic, high-performance use by partners and end users, a critical factor given that these kinds of significant updates result in an increase in harvesting activities and use. The approach used in generating and adding the linked data allows for regular updates to the decoration without significant timing or technical challenges. It is likely that the markup may evolve over the coming months so this release should be considered experimental and subject to change.

### Making Library Data Relevant on the Web

The Schema.org activity and associated vocabularies offer a clarified middle ground between rich, very diverse domains on the Web where context does not exist. Web intermediaries like Google, Bing, and Yahoo! focus on interpreting web users’ needs and connecting them to the most appropriate web resources. Using linked data and the Schema.org vocabularies as a starting point, rich domains like libraries, retailers, publishers, governments, and scientists can surface in this webscale interpretation with more context and clearer intent.

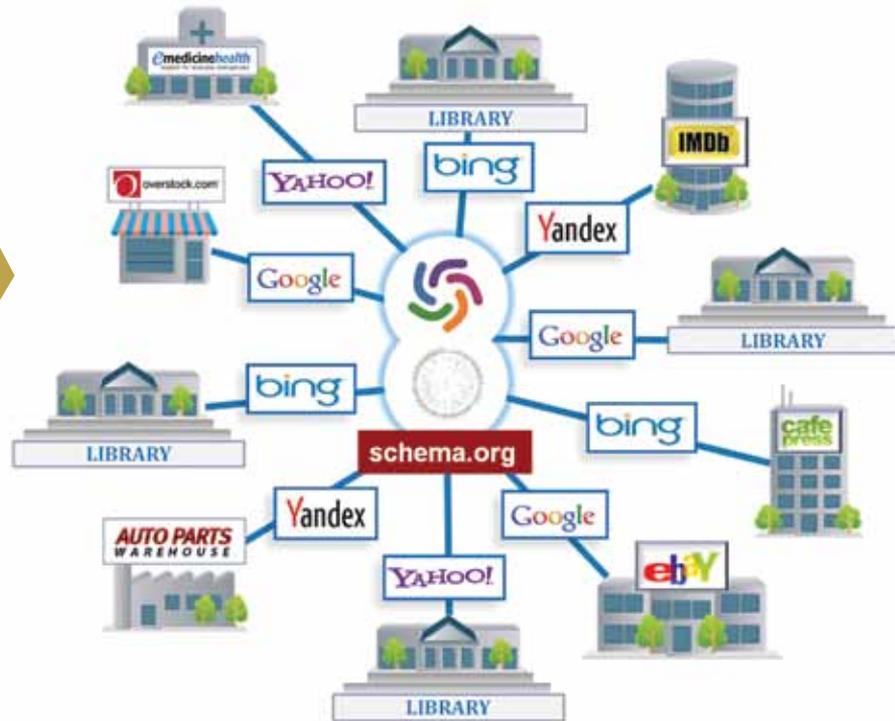
Webscale means three things in this exercise for OCLC:

- 1 **A large volume of data**  
Rather than experiment with a subset, apply the markup decoration to more than 250 million bibliographic items in WorldCat.org. The initial decoration included Virtual International Authority File (VIAF), Faceted Application of Subject Terminology (FAST), Library of Congress Authorities, and Dewey.
- 2 **Quick, large-scale iterations**  
The design of the technical and data infrastructures allow for quick iterations and updates to the data. No one “right” way to

CONTINUED »

## WORLDCAT.ORG & SCHEMA.ORG

Using linked data and the Schema.org vocabularies as a starting point, rich domains like libraries, retailers, publishers, governments, and scientists can surface in this webscale interpretation with more context and clearer intent.



do this markup exists; therefore, additional vocabularies will be identified, better expressions clarified, and more meaningful connections made over time. OCLC's data infrastructure and the architecture of WorldCat.org both support this kind of iteration and exposure for the large dataset and associated decorations.

**3 Ongoing, Open Discussions & Community-Based Learning**  
Participation in the Schema.org work reflects contemporary web expectations of trying many things to get to a better place overall. Interested parties can learn more about engaging with OCLC, and look at the proposed library vocabulary extension at the *Linked data at OCLC* webpage. See relevant links below.

### The Future

Like all experiments, this project is the basis for further iterations of work, based upon results, to further enhance linked data capabilities in WorldCat data. This work falls into several categories.

#### Vocabulary:

As previously discussed, the exposure of WorldCat data was approached from the viewpoint of the consumer not familiar with libraries. For a search engine company or a general web consumer, the vocabulary most likely to be generally accepted is Schema.org, (as its markup is already found on some seven

percent of pages crawled by Google and Bing). However, as recognized by the development of a library ontology to supplement schema.org markup for WorldCat data, any vocabulary will need to be extended to address the lack of some details.

#### The library ontology:

This is designed as a conversation starter for a recommended extension to Schema.org and not a complete ontology. This conversation should be encouraged and pursued with other organizations and individuals in the library and Semantic Web domains. If a consensus can be formed around this proposal, there is a good chance that the W3C-backed group behind Schema.org will accept it. If accepted, it will benefit everyone on the Web by providing structured library data. Libraries will benefit by being able to more broadly share information about their resources.

#### Access to data:

RDFa embedded in HTML is only one way of providing access to WorldCat as linked data. The use of content negotiation to deliver this data as RDF, in formats such as JSON, RDF/XML, Turtle, etc., is one way to investigate the delivery of this data. Scraping the RDF from the content of a WorldCat webpage, although powerful, is not the ideal access method for all circumstances. In the coming months the best ways to provide

access to this data will be explored; this will include talking to potential data consumers and identifying services that can be provided around it.

#### Productization:

By definition this experiment is designed to improve the core part of the production service that looks after WorldCat. As the ways of describing and providing access to WorldCat linked data evolve, work to enhance the OCLC infrastructure to implement this, as part of normal processes, will occur. With the size and uses of WorldCat, and the number of processes in place to add and maintain data within it, this is not an insignificant task, but the potential benefits make it one worth undertaking. As things iterate from this experiment, there will be as much work behind the scenes as will be visible on the surface.

#### A linked view of the world:

There have been implicit linkages held in WorldCat data for years, as demonstrated by links to VIAF, FAST, Dewey, Library of Congress and other authoritative resources that have been surfaced by this experiment. Making these links explicit, identifiable, and accessible will open up potential for new services and new ways of thinking about the process of creating, managing, and sharing data. Work is underway to identify other links that could be exposed to more authoritative sources. Suggestions for more links and ways to map to them are encouraged.

As other contributions in this *ISQ* issue indicate, linked data is entering the vocabulary, and practice, of many in the metadata community. This experiment represents a strong commitment from OCLC toward the debate around the



Making these links explicit, identifiable, and accessible will open up potential for new services and new ways of thinking about the process of creating, managing, and sharing data.

best ways forward and the potential benefits of linked data. Join the conversation, as we iterate forward from this initial experimental step. If you have questions or comments about what we have done or might do next, please contact us at [data@oclc.org](mailto:data@oclc.org).

IIP | doi:10.3789/isqv24n2-3.2012.05

**TED FON** ([fonst@oclc.org](mailto:fonst@oclc.org)) is Executive Director, Data Services & WorldCat Quality at OCLC. **JEFF PENKA** ([penkaj@oclc.org](mailto:penkaj@oclc.org)) is Director/Global Product Manager, QuestionPoint Services at OCLC. **RICHARD WALLIS** ([richard.wallis@oclc.org](mailto:richard.wallis@oclc.org)) is Technology Evangelist in OCLC's Birmingham, UK office.

**Apache™ Hadoop™**  
[hadoop.apache.org/](http://hadoop.apache.org/)

**Dewey.info**  
[dewey.info/](http://dewey.info/)

**Experimental “library” extension vocabulary for use with Schema.org**  
[purl.org/library/](http://purl.org/library/)

**Faceted Application of Subject Terminology (FAST)**  
[www.oclc.org/research/activities/fast/](http://www.oclc.org/research/activities/fast/)

**Library of Congress Authorities**  
[authorities.loc.gov/](http://authorities.loc.gov/)

**Linked data at OCLC**  
[www.oclc.org/data.html](http://www.oclc.org/data.html)

**Open Data Commons license (ODC-BY)**  
[opendatacommons.org/licenses/by/](http://opendatacommons.org/licenses/by/)

**Schema.org**  
<http://schema.org/>

**Virtual International Authority File (VIAF)**  
[www.oclc.org/viaf/](http://www.oclc.org/viaf/)

**W3C Semantic Web Linked Data Specifications**  
[www.w3.org/standards/semanticweb/data](http://www.w3.org/standards/semanticweb/data)

**WorldCat®**  
[www.worldcat.org](http://www.worldcat.org)



RELEVANT  
LINKS


 Antoine  
Isaac

 Robina  
Clayphan

 Bernhard  
Haslhofer

# EUROPEANA:

## Moving to Linked Open Data

ANTOINE ISAAC, ROBINA CLAYPHAN, AND  
BERNHARD HASLHOFER

**EUROPEANA** is the European Union's flagship digital cultural heritage initiative. The Europeana portal, launched in November 2008, showcases the possibility of cross-cultural domain interoperability on a pan-european level. To date, metadata and thumbnails for over 23 million objects have been aggregated from over 1500 providers from the library, archive, museum, and audiovisual domains. Offering simple and advanced search functionality, or browsing using various parameters, users can link from the representations of the objects held in the portal to the source objects held at the provider institutions.

From the outset Europeana was conceived as more than just a huge data repository fronted by a portal application. It was hoped that cultural heritage communities were ready to think outside the traditional information silos and adopt a linked data paradigm that would enable the development of shared semantic context. Linked Data is a data publishing technique that uses common web technologies to connect related data and make them accessible on the Web. Linked Open Data implies that reuse restrictions have been removed from the metadata. Moving to such a model may mean that in future the portal is seen as the reference application of Europeana but that its main function is that of a rich data service that allows third parties to take the data freely and reuse it to create new knowledge and applications.

This is an ambitious goal, requiring a change of

perspective on the part of the guardians of cultural heritage resources: "This mentality shift is a big leap, since it requires cultural heritage institutions to think, not primarily within the boundaries of their particular collections, but in terms of what these collections might add to a bigger, complex and distributed information continuum coupled with various contextual resources." [Concordia] Every aspect of this ambition offers major challenges: technical, legal, policy level, linguistic, financial, etc. All this has meant that a fully linked open data position could only be achieved in an iterative fashion and by keeping providers involved at all times.

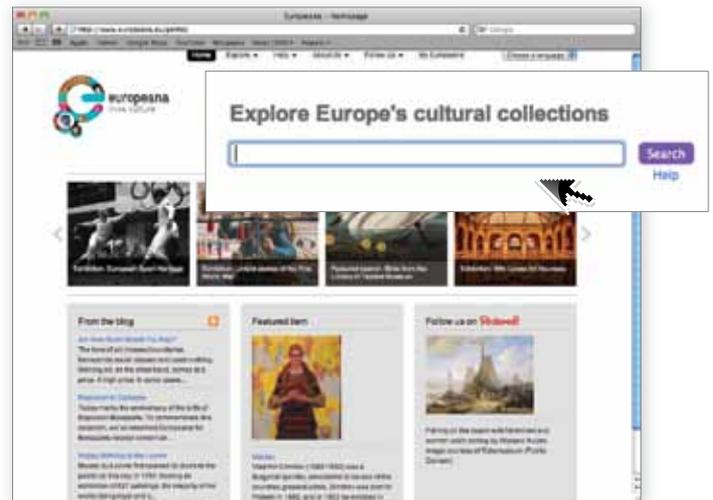
As a technical starting point, Europeana carried out a linked open data pilot project. The remainder of this paper gives an outline of the processes of the pilot project and areas for future work.

## From local data standards to the Europeana Semantic Element set

The shift required to get from individual data silos with curated data to the distributed information continuum enabled by linked open data was not going to be accomplished in one giant leap for mankind. Providers use many different metadata formats ranging from well-known, sophisticated, internationally-maintained standards to home-grown formats that had evolved in one institution over time. Coming from different cultural heritage sectors, these standards also embody different views of the resources they curate: for example, event-based versus object-centric descriptions. The first step was therefore to demonstrate the possibility of interoperability between the silos and, in parallel, to start examining the many other aspects already mentioned. Not the least of these, but not covered in this paper, was to develop trust between partners and build a community with a common understanding of the aims of the enterprise, including the key concept of “open data.”

To achieve some measure of data interoperability, a common dataset was defined to which all participating providers could map a reasonably useful set of metadata. This initial metadata schema is the Europeana Semantic Elements (ESE). This is essentially a Dublin Core application profile: an element set based on a subset of the Dublin core elements with several Europeana-specific fields added to support specific portal functionality. The documentation giving information on ESE can be found in the “Technical requirements” section of the Europeana website. This schema is the current metadata format used in the European production system.

As a basic solution to Europeana interoperability problems, ESE suffers from many issues. First, it is a “flat” model that aggregates in one and the same record metadata fields that can apply to different entities. This breaks the “one-to-one” principle and causes great confusion; for example, some providers use rights- or date- related fields to give information for the “real-world” resources they hold, while others use the same fields for data about the digital resource that represents these items. Significantly for a linked data approach, most of the data provided contains simple string values for the metadata fields. Linked data depends on resources being identified with (HTTP) Uniform Resource Identifiers (URIs) in order to create the links. Simple string values prevent properly linking an item ingested by Europeana to other objects (e.g., a series of portraits), or to contextual entities as represented by complex



Moving to such a model may mean that in the future the portal is seen as the reference application of Europeana but that its main function is that of a rich data service that allows third parties to take the data freely and reuse it to create new knowledge and applications.

resources, e.g., a creator with many name variations, or a broader concept that is part of an online thesaurus, all of which could help improving access to Europeana items.

## From ESE to the Europeana Data model (EDM)

The RDF-based Europeana Data Model (EDM) was developed by the Europeana community as an alternative to the ESE schema and aimed at solving the shortcomings mentioned. The development process took full account of Europeana's firm belief in the benefits of Semantic Web and Linked Data technology for the culture sector, which have been articulated in the reports of the W3C Library Linked Data Incubator Group. It is a more flexible and precise model than ESE which offers the opportunity to attach every statement to the specific resource it applies to and also reflects some basic form of data provenance.

CONTINUED »

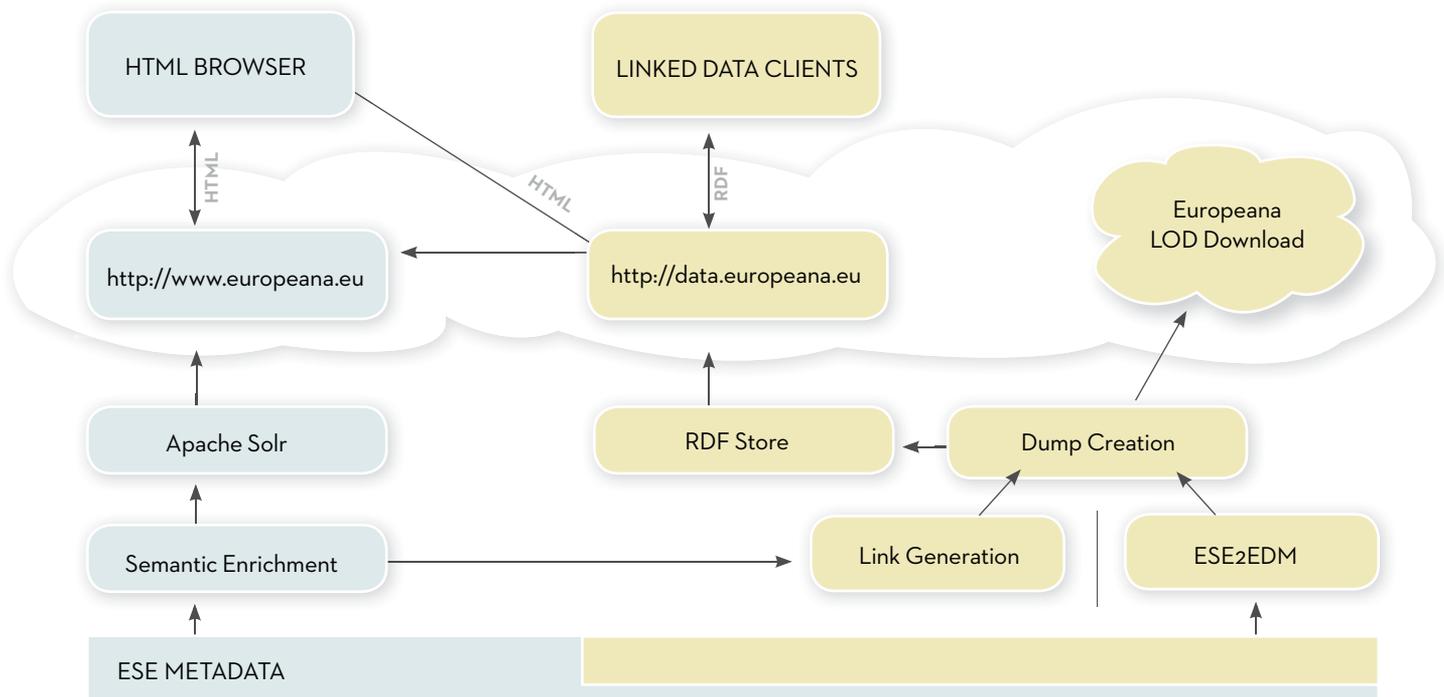


Figure 1: data.europeana.eu - Technical Architecture

The main requirements identified for the development of EDM included:

- » Distinguish between a “provided item” (painting, book) and digital representations.
- » Distinguish between an item and the metadata record describing it.
- » Allow ingesting multiple records for the same item, containing potentially contradictory statements about it.
- » Provide support for contextual resources, including concepts from controlled vocabularies.

By providing the mechanism to distinguish all these aspects of a resource, EDM allows the representation of different perspectives on a given cultural object. It also enables the representation of complex—especially hierarchically structured—objects, as in the archive or library domains. Finally, it allows the representation of contextual information, in the form of entities (places, agents, time periods) explicitly represented in the data and connected to a cultural object.

Rather than systematically introducing new elements, EDM reuses and links to existing reference vocabulary elements, such as the Open Archives Object Reuse and Exchange Model (OAI-ORE), Dublin Core, and the W3C SKOS model for Knowledge Organization Systems. These various features are fully described in the EDM Primer on the Europeana Professional website.

### The Linked Data Pilot - data.europeana.eu

As mentioned earlier, many issues stood in the way of the immediate adoption of Linked Open Data (LOD) in the Europeana production system:

- ❶ Lack of metadata expressed in EDM
- ❷ Missing links to other sources
- ❸ The absence of data provider agreements explicitly permitting the release of the metadata into the public domain.

As a proof of concept, a Europeana Linked Data Pilot was built at data.europeana.eu. It is technically de-coupled from the Europeana production system and allows those data providers, who want to make their data available as Linked Open Data, to opt for their metadata to be openly published on the Web.

The overall approach is shown in Figure 1 and an outline description follows. A fuller technical description can be found in the paper produced by Haslhofer and Isaac in 2011.

- ❶ Extract the subset of ESE XML metadata that had been submitted by the providers who had expressed the wish to become part of the pilot.
- ❷ Convert the ESE data to EDM using the mapping that had been defined. This mapping also covered the creation of the EDM entities (items, aggregations, proxies), the assignment of dereferencable HTTP URI identifiers to these entities, as well as the attachment of the relevant metadata fields

to each new entity. The mapping between ESE and EDM is implemented in an XML stylesheet. The result is an RDF/XML representation of each data provider's metadata.

- 3 Two strategies are followed for linking data.europeana.eu resources with other web resources:
  - » Semantic enrichment data that is created by Europeana, after it has ingested metadata from its data providers, is fetched. This data consists of links to four types of reference resources: GeoNames for places, GEMET for general topics, the Semium time ontology for time periods, and DBpedia for persons, currently generating over four million links. Since the enrichments are links, they perfectly fit EDM and the Linked Data approach.
  - » A simple ad-hoc linking strategy whereby existing resource identifiers that are part of the metadata are used to create links to other Linked Open Data services that hold information about objects that are also served by data. For the time being, this only concerns the Swedish cultural heritage aggregator (SOCH).

Data dumps were generated from the resulting RDF/XML files together with the supplied/generated links. These are then made available as dump files and also ingested into an RDF store. Incoming HTTP requests are answered either by the RDF store (if they have an RDF-specific Internet media type in the HTTP Accept header field) or redirected to the Europeana portal (for standard HTML requests).

### EDM modeling patterns

Figure 2 shows the basic structures of EDM networked resources after the flat ESE data is transformed into EDM for the linked data pilot.

The following sections explain each resource further and indicate the properties that should be attached to their instances.

#### Item (Provided Cultural Heritage Object)

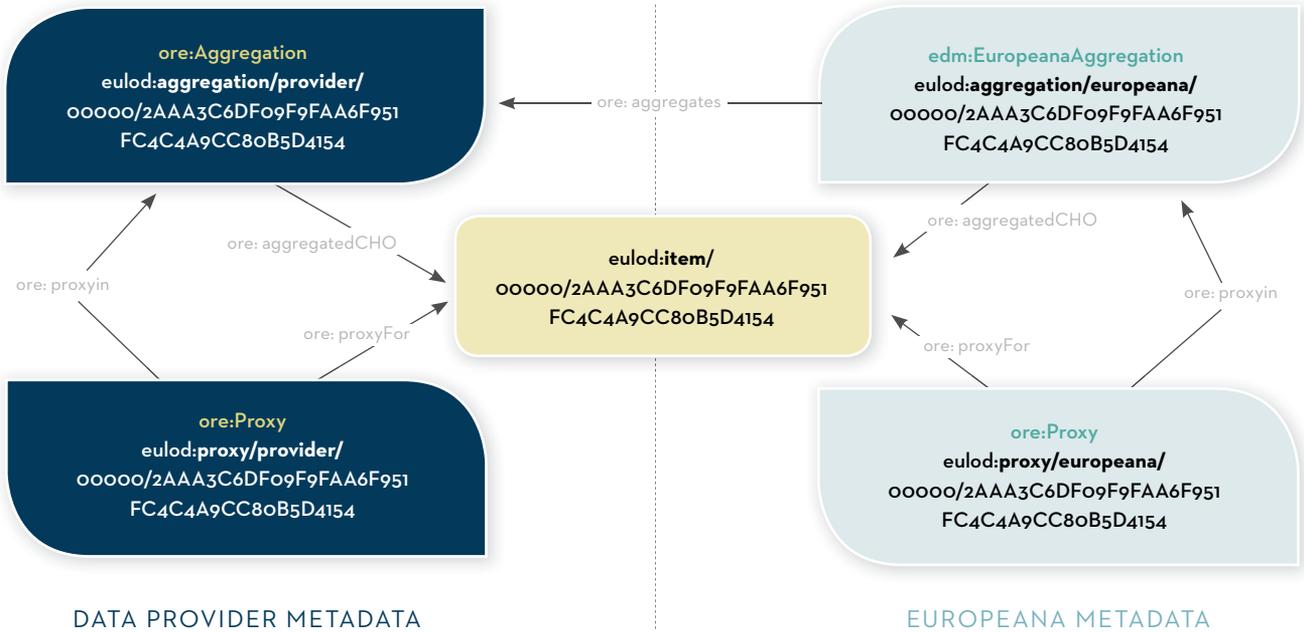
Item resources (typed as Provided Cultural Heritage Object (CHO)) represent objects (painting, book, etc.) for which institutions provide representations to be accessed through Europeana. Provided CHO URIs are the main entry points in data.europeana.eu. A Provided CHO is the hub of the network of relevant resources and, when applicable, will link out to other linked data resources about the same object via owl:sameAs statements. In the pilot, no descriptive metadata (creator, subject, etc.) is directly attached to object URIs. It is instead attached to the proxies that represent a view of the object, from a specific institution's perspective (either a provider or Europeana itself).

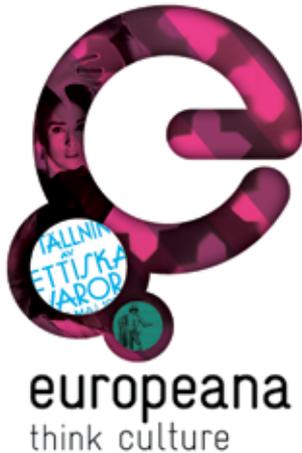
#### Provider's proxy

Proxies originate from the OAI-ORE model and are used to separate the item itself from the descriptive statements (creator, subject, date, etc., mostly coming from ESE's Dublin Core fields) for the item, which are contributed by a provider. They enable the separation of different views of the same item

CONTINUED »

Figure 2: Basic structure of EDM networked resources





Europeana's mission includes becoming a trusted source of information, while encouraging more open data circulation in the culture sector. To this end, provenance and licensing information are crucial—whether about the cultural items being accessed or about the metadata on these items.

that may be the focus of multiple aggregations from different providers. In every case, there will be one proxy for the provider descriptive data for an item and another for the data created by Europeana.

#### ***Provider's aggregation***

These resources provide data related to a provider's gathering of digitized representations and descriptive metadata for an item. They are related to digital resources about the item, be they files directly representing it or webpages showing the object in context. They may also provide controlled rights information applying to these resources. Finally, provenance data is given in statements using the specific EDM properties.

#### ***Europeana's proxy***

Europeana proxies are the second type of proxies served at data.europeana.eu. They provide access to the metadata created by Europeana for a given item, distinct from the original metadata from the provider. Here, one can find statements indicating a normalized date associated with the object. These proxies also have statements that link them to places, concepts, persons, and periods from external datasets, as mentioned earlier.

#### ***Europeana's aggregation***

A Europeana aggregation bundles together the result of all data creation and aggregation efforts for a given item—the provider's and Europeana's own. It aggregates the provider's aggregation, which in turn will connect to the provider's proxy. Not shown in the diagram, but linked to the provider aggregation, are the digitized resources europeana.eu serves for the item.

### **Issues and future work**

#### ***Achieving fully open data***

When the results of the linked data pilot were first launched in June 2011, it contained 3.5 million objects taken from the datasets of volunteer institutions. These could not be released under fully open terms due to the evolving understanding of the Data Exchange Agreement. In February 2012, a second version of data.europeana.eu was released that, although still a pilot, now contains fully open metadata (CC0 – public domain dedication). It has a smaller but still substantial subset of cultural heritage data, at 2.4 million objects, but this must be seen in context: the qualitative step of having fully open publication is crucial to Europeana and forms the basis of an active advocacy campaign to persuade more of the community to open their data for the benefit of end users. In order to make the message more accessible to the public, an animation explaining the connection between linked data technology and open data policies has been released. A virtuous circle is envisaged in which third parties use the open data to develop innovative applications and services, which in turn stimulates end users' interest in digitized heritage, and this, in turn, demonstrates to cultural heritage institutions the value of releasing more open data.

#### ***Improving connectivity of the data***

Source data in Europeana is of varying degrees of richness and is all mapped to ESE, which is based on simple text string values. While achieving interoperability, this often entails losing some of the richness of the more detailed formats. In particular, it means any provider that has used contextual resources (authority files, thesauri, etc.) will have lost those relationships. In the context of the linked

The transition from Europeana URIs to dereferencable HTTP URIs for EDM aggregations and proxies was a major challenge in the conversion process. The main Europeana production system and the Europeana Linked Open Data Prototype are still two distinct systems so a bridge was needed between the identification mechanisms in place.



data pilot this means that internal connectivity is very low. Linkage exists between the provided CHO—aggregation—proxy resources that come with the EDM model, but no “semantic” links between the items or the proxies that represent them. Ideally, many provider contextual resources could be fed into Europeana together with the object metadata and provide internal links. This includes, among others, concepts from shared domain thesauri or place resources, which are already used in the description for different objects in a collection or even across collections. Publishing the data together with its companion thesaurus and authority file has already been demonstrated in the Amsterdam Museum Linked Open Data prototype. Europeana is currently working on this and there are case studies on the Europeana Professional website that show how it can be done.

For achieving external connectivity, Europeana’s enrichment process is used and this generates semantic links from specific fields in the ESE data. Because it has to deal with very heterogeneous collections, Europeana is bound, for the moment, to use simple data enrichment techniques despite the associated errors. Improving the way the creation of enrichment values is handled would improve this situation, however. In addition, alignments will be extended to other resources such as the Virtual International Authority File (VIAF) and other relevant initiatives from the community.

#### **Disseminating meta-metadata**

An important Europeana requirement is to communicate meta-level information about the data it publishes. Europeana’s mission includes becoming a trusted source of information, while encouraging more open data circulation in the culture sector. To this end, provenance and licensing information are crucial—whether about the cultural items being accessed or about the metadata on these items.

Linked Data technology still lacks a fully standardized suite to express such meta-level information so various existing solutions (based on OAI-ORE resource maps) were combined to supply the required licensing and provenance data for the present. This choice is similar to other institutions, for example, the New York Times’ linked data service. Relevant ongoing efforts (the W3C Provenance Working Group and DCMI’s Provenance Task Group) are being followed in the hope of adopting a fully consensual approach in the future.

#### **Complexity and navigability**

The requirements for EDM to distinguish different data sources and apply the data precisely to different resources results in the creation of complex networks of aggregations, proxies, and other resources. This has many benefits but it also raises the barrier to data access and consumption. As well as adding extra complexity to the RDF graphs published, the proxy pattern is counter-intuitive for linked data practitioners. It causes confusion in particular when finding statements about something (for example, a painting) that are attached to a resource that is not, strictly speaking, standing for that painting (i.e., the proxy). The temptation was to simplify the task for the linked data consumers by duplicating the statements attached to the proxies onto the “main” resource for the provided item thereby allowing direct access to the statements, i.e., not mediated through proxies. Although this was not done, it may still happen in response to feedback from data consumers. In the longer term, it is hoped that W3C will standardize named graphs for RDF thereby allowing EDM to meet the requirement to track provenance without the need for proxies.

#### **HTTP URI design**

The transition from Europeana URIs to dereferencable HTTP URIs for EDM aggregations and proxies was a major challenge in the conversion process. The main Europeana

CONTINUED »

production system and the Europeana Linked Open Data Prototype are still two distinct systems so a bridge was needed between the identification mechanisms in place. Europeana's local identifiers were therefore used for the dereferencable URIs in the Europeana LOD prototype. This resulted in persistence difficulties when collections were reharvested. Both the LOD infrastructure and the underlying Europeana identification mechanism will have to find better strategies in the future.

### Integration with other data

A further area of future work will be the compatibility between Europeana data and other initiatives that promote the availability of structured metadata on the Web, such as schema.org. This should increase the visibility of Europeana on the Web.

### Conclusion

Data modeling and description practices differ across the cultural heritage sector, varying in levels of granularity, focus of interest, use of standards, and application of vocabularies. It was important that the solution chosen by Europeana should reuse existing standards and be flexible enough in its approach to interoperability to allow their co-existence with custom ones from across the sector. Because Europeana wants to reuse and be reused, a web-based open technology was ideal to make it simple to connect data together and share it. Such semantic web and linked data technologies directly relate to open data strategies.

The Europeana linked data pilot produced a body of open metadata represented in the EDM. This allows the representation of different perspectives and basic provenance information on any given cultural object. It is anticipated that future data.europeana.eu dataset releases will reflect the lessons learned with respect to the model's complexity, dealing with provenance and increasing Europeana's internal and external connectivity. Key contributions will include applying more semantic enrichment techniques and aggregating richer EDM metadata from data providers instead of flat ESE records.

Europeana is a strong advocate of the benefits of semantic web and linked data technology for the culture sector and the associated opportunities for opening data for imaginative reuse by third party developers and end users. By developing its data model based on these principles and producing this pilot set of data, the foundations are in place for building a shared semantic context for cultural heritage data. | IP | doi: 10.3789/isqv24n2-3.2012.06

---

**ANTOINE ISAAC** (aisaac@few.vu.nl) is Scientific Coordinator at Europeana, The Hague, The Netherlands. **ROBINA CLAYPHAN** (robina.clayphan@kb.nl) is Interoperability Manager at Europeana Foundation. **BERNHARD HASLHOFER** (bernhard.haslhofer@cornell.edu) is Fellow Postdoc Associate (Marie Curie) at Cornell University.



## RELEVANT LINKS

**Amsterdam Museum in Europeana Data Model RDF**  
semanticweb.cs.vu.nl/lod/am

**Animation about Linked Open Data by Europeana**  
vimeo.com/36752317

**Concordia, Cesare, Stefan Gradmann, and Sjoerd Siebinga.**  
*Not just another portal, not just another digital library: A portrait of Europeana as an application program interface.*  
*IFLA Journal*, 2010, 36: 61  
DOI: 10.1177/0340035209360764 <http://ifl.sagepub.com/content/36/1/61>

**DBpedia**  
dbpedia.org/

**DCMI Metadata Provenance Task Group**  
dublincore.org/groups/provenance/

**Dublin Core Element Set**  
dublincore.org/documents/dcmi-terms/

**EDM Case Studies**  
pro.europeana.eu/case-studies-edm

**EioNet GEMET Thesaurus**  
www.eionet.europa.eu/gemet/

**Europeana Data Model (EDM) Documentation**  
pro.europeana.eu/edm-documentation

**Europeana Semantic Elements Technical Requirements**  
pro.europeana.eu/technical-requirements

**Geonames**  
geonames.org

**Haslhofer, Bernhard, and Antoine Isaac.** *data.europeana.eu: The Europeana Linked Open Data Pilot. Proceedings International Conference on Dublin Core and Metadata Applications (DC-2011). The Hague, Netherlands, September 21-23, 2011.*  
[dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/55/14](http://dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/55/14)

**Open Archives Object Reuse and Exchange Model (OAI-ORE)**  
[www.openarchives.org/ore/](http://www.openarchives.org/ore/)

**Semium.org**  
semium.org

**SKOS Simple Knowledge Organizational System (W3C)**  
[www.w3.org/2004/02/skos](http://www.w3.org/2004/02/skos)

**The New York Times Linked Open Data<sup>Beta</sup>**  
data.nytimes.com

**W3C Library Linked Data Incubator Group**  
[www.w3.org/2005/Incubator/lld/](http://www.w3.org/2005/Incubator/lld/)

**W3C Provenance Working Group**  
[www.w3.org/2011/prov/](http://www.w3.org/2011/prov/)

**W3C RDF Working Group**  
[www.w3.org/2011/rdf-wg/](http://www.w3.org/2011/rdf-wg/)

A judgement formed about something;  
a personal view, attitude, or appraisal



Jon Voss

JON VOSS

## LODLAM State of Affairs

In 2009, a group of developers, librarians, archivists, historians and technologists, myself included, had the idea of working with disparate datasets from a series of libraries, archives, and museums to enable new understanding of the American Civil War. We sought to use Linked Data to make connections between the various datasets. In other words, we sought to simply break the information down into a format that would allow us to discover, describe, and navigate the relationships between the various datasets, the assets represented by metadata, and basic organizing units of the Civil War (like regiments, battles, etc.).

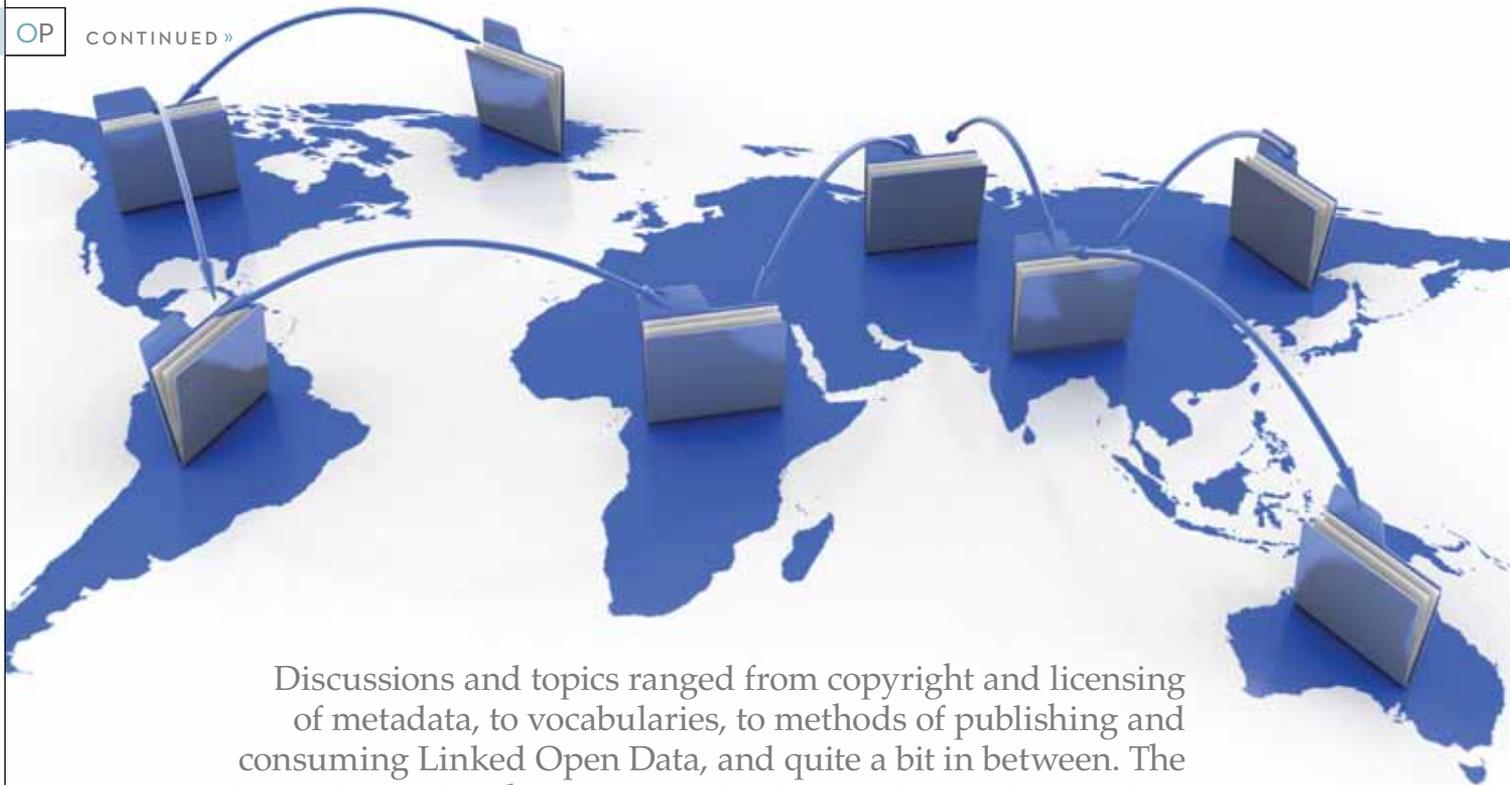
Of course, we were a ragtag group applying for a digital humanities grant, and the review panel was not sure what to make of us. What's more, we found that even though we thought we did a pretty good job of describing what Linked Open Data was, people either loved the idea or didn't get it at all. And the people who loved the idea weren't able to describe it to the people who didn't get it. So maybe they just weren't ready for us and the project went unfunded. But what we found was that there was an appetite for educating the broader community of libraries, archives, and museums about the concepts of Linked Open Data, and we were funded to help do it. So we joined forces with colleagues around the world who were interested and created #lodlam and the Linked Open Data in Libraries, Archives, and Museums Summit, which took place June 2-3, 2011.



Perhaps this sounds all too familiar. As you know in the international standards community, this isn't a new idea, though it may be an idea whose time has come. Interest in the topic has grown exponentially, even if adoption of Linked Data technologies has not necessarily followed suit. Yet the combination of legal tools such as Creative Commons licenses, the increasingly common publishing of open datasets by institutions, and the growing number of Linked Data examples in the library world suggest that a rather monumental shift may be afoot.

Last year, we had hoped to gather about 50 people for the Summit, catalysts in their fields, but were overwhelmed with the interest and so expanded to accommodate 100 participants. The participants set their own agenda and pursued a diverse range of topics over the two day meeting.

CONTINUED »



Discussions and topics ranged from copyright and licensing of metadata, to vocabularies, to methods of publishing and consuming Linked Open Data, and quite a bit in between. The Summit certainly helped galvanize international collaboration, and led to regional meet-ups around the world.

In total, 85 organizations from 17 countries were represented. Delegates included developers, scholars, researchers, policy makers, funders, and vendors from across the humanities and sciences. Discussions and topics ranged from copyright and licensing of metadata, to vocabularies, to methods of publishing and consuming Linked Open Data, and quite a bit in between. The Summit certainly helped galvanize international collaboration, and led to regional meet-ups around the world.

Consider that in the last year alone we've seen Linked Open Data projects pertinent to bibliographic data from Stanford University, the National Library of Spain, the British Library, and the National Library of Germany. Europeana has published metadata on 2.4 million items gathered from over 200 institutions as Linked Open Data, which will soon increase to 15 million. Schema.org was launched by a collaboration of the world's biggest search engines. The W3C Library Linked Data Incubator Group issued their final report with key recommendations for libraries. And the Library of Congress announced that they would move from MARC to a Linked Data model.

As we continue to discuss the issues at conferences around the world, curators, librarians, technologists, and vendors are digging deeper into the questions of Linked Data and the Semantic Web. Small, domain-specific test cases are beginning to get off the ground. The foundations of commonality and collaboration so long supported by the international standards

community are providing significant building blocks for the Web of Data. In 2013, we'll be gathering again for another International Linked Open Data in Libraries, Archives, and Museums Summit, and look forward to your continued representation in discussions and participation in #lodlam projects. The next Summit will be held in Montreal June 19-20, 2013. | [IOP I](#) doi: 10.3789/isqv24n2-3.2012.07

**JON VOSS** ([jon.voss@wearewhatwedo.org](mailto:jon.voss@wearewhatwedo.org)) is Historypin Strategic Partnerships Director at We Are What We Do and Chair of the Organizing Committee for International Linked Open Data in Libraries, Archives, and Museums Summit.

**LOD-LAM: The International Linked Open Data in Libraries, Archives, and Museums Summit**  
[lodlam.net/](http://lodlam.net/)



**RELEVANT  
LINKS**



Thomas  
Elliott



Sebastian  
Heath



John  
Muccigrosso

THOMAS ELLIOTT, SEBASTIAN HEATH, AND JOHN MUCCIGROSSO

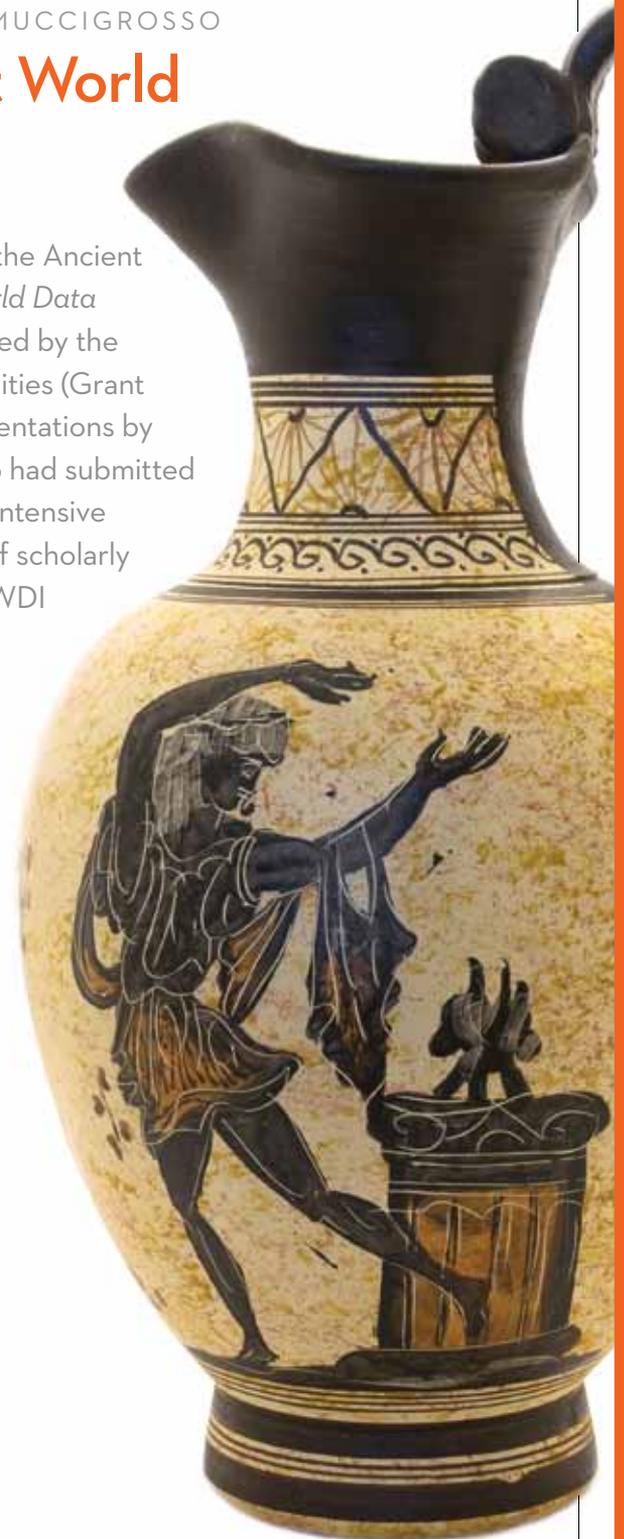
## Report on the Linked Ancient World Data Institute

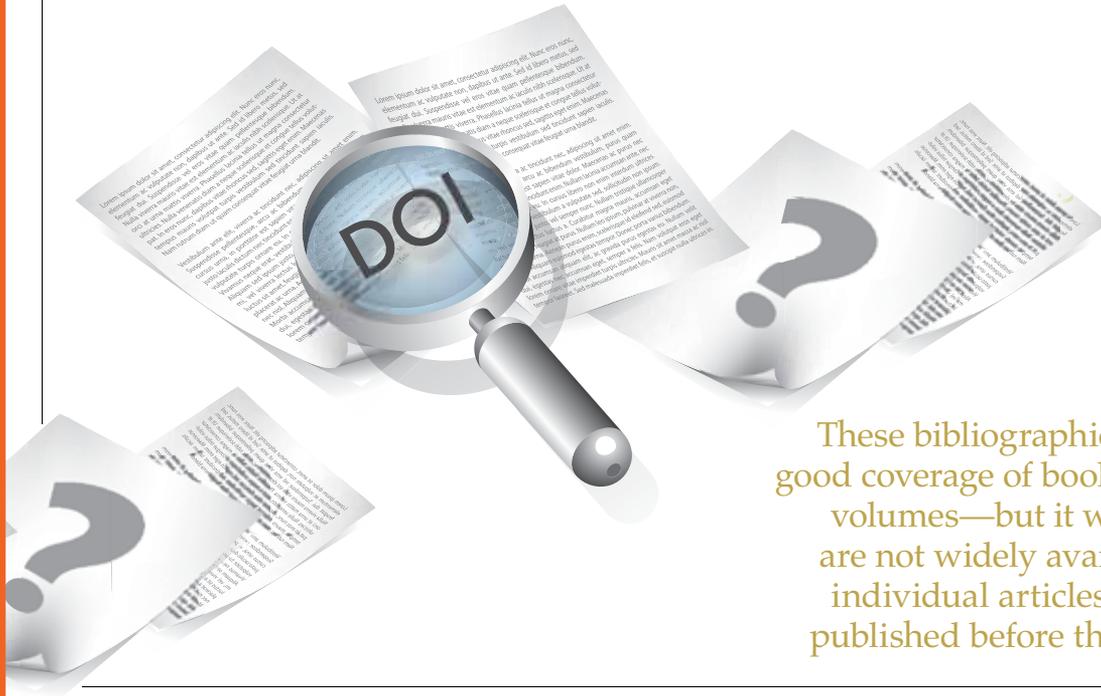
From May 31 to June 2nd, 2012, the Institute for the Study of the Ancient World at New York University hosted the *Linked Ancient World Data Institute* (LAWDI), an internationally attended workshop funded by the National Endowment for Humanities' Office of Digital Humanities (Grant number: HT5004811). This three-day event mixed longer presentations by invited speakers with presentations by twenty applicants who had submitted statements of interest on why their work would benefit from intensive interaction with colleagues also pursuing digital publication of scholarly resources on the public internet. This was the first of two LAWDI sessions, with the second to be held at Drew University from May 30 to June 1, 2013.

LAWDI's intellectual scope is the ancient Greek and Roman Mediterranean and the Ancient Near East. Taken as both geographic and chronological markers, these terms encompass modern academic disciplines that have long histories of creating digital resources, many of which are already accessible via HTML-based websites. And while it is too optimistic to say that these disciplines have always maximized the potential of interdisciplinary work, there is a continuity of cultural development and a degree of contact that gives many commonalities to the study of the early civilizations of Mesopotamia, the subsequent eras of Greek and Roman cultural prominence, and the ongoing reworking of ancient precedent by later Byzantine and Syriac societies. Accordingly, one premise of LAWDI is that publication of well-structured and reusable digital resources will benefit all scholars, as well as the interested public who are working in these areas.

In addition to being academically inclusive, LAWDI also took a very open approach to the concept of Linked Open Data. Most of the attendees at LAWDI were experts in the content and methods of their respective disciplines, rather than in the technical details of web architecture. Additionally, many present were also included museum professionals, librarians, and archivists who curate so-called ancient world data, such as archaeological fieldwork archives and bibliographic resources. Again, there was no expectation that participants came with experience in implementing

CONTINUED »





These bibliographic resources provide good coverage of books—that is physical volumes—but it was noted that there are not widely available identifiers for individual articles, particularly those published before the adoption of DOIs.

Linked Open Data so that the organizers recognized that a three-day workshop was not enough time to develop complete technical proficiency.

Accordingly, LAWDI began with a focus on two aspects of current best practices.

- 1 Early sessions stressed the importance of establishing stable URIs that allow fine-grained access to scholarly resources. Examples of current work included the URIs that Pleiades is establishing for ancient geographical entities and URIs for numismatic concepts established by Nomisma.org.
- 2 Presenters stressed that progress going forward depends on high-quality, automatically parsable data being available when those URIs are de-referenced. Of course, RDF can play a role here, and attendees were introduced to basic concepts as “triples” and “things, not strings.” But there was also discussion of RDFa, JSON, KML, and Atom as reasonable formats that allow machine-based reuse of ancient world data. As an example of such reuse, more than one presentation discussed the Pelagios Project, which is aggregating references to Pleiades URIs via the Open Annotation Collaboration RDF vocabulary—and is currently one of the best examples of the potential for Linked Open Data to enable new forms of discovery of scholarly resources. In particular, the overlap between geographic named entities and discovery of ancient textual sources that refer to those entities is being pushed forward by the participation of the Perseus Digital Library and Google Ancient Places in the Pelagius consortium.

While participant presentations were spread throughout the three-day program, it quickly became apparent that many of the applicants came to LAWDI with very basic questions, all of which were very welcome. In general, those of us working on the digitization of the Ancient World recognize the importance of reusing existing vocabularies. But we also recognize that it is very easy to push the limits of what generic vocabularies such as the Dublin Core allow us to communicate. For example, does `dcterms:creator` refer to the webpage being de-referenced at a URI or to the ancient artist who created the object being described by the document found there? Such a question is recognizable as falling under the rubric of “HTTP Issue 14”—now open as “Issue57: Mechanisms for obtaining information about the meaning of a given URI”—and the invited speakers strove to highlight such issues and to illustrate both the “fragment identifier” and “303 redirect” mechanisms for solving them. It is not clear that the expertise to implement such solutions is widely available in the academic computing environments that were typically available to LAWDI participants.

Discussion during the workshop also highlighted the availability of bibliographic linked data as a pressing need for scholarly initiatives. The Library of Congress’ `id.loc.gov` server, OCLC’s VIAF service, and the resources OCLC makes available via WorldCat were all highlighted. These resources provide good coverage of books—that is physical volumes—but it was noted that there are not widely available identifiers for individual articles, particularly those published before the adoption of DOIs. While JSTOR does provide retrospective

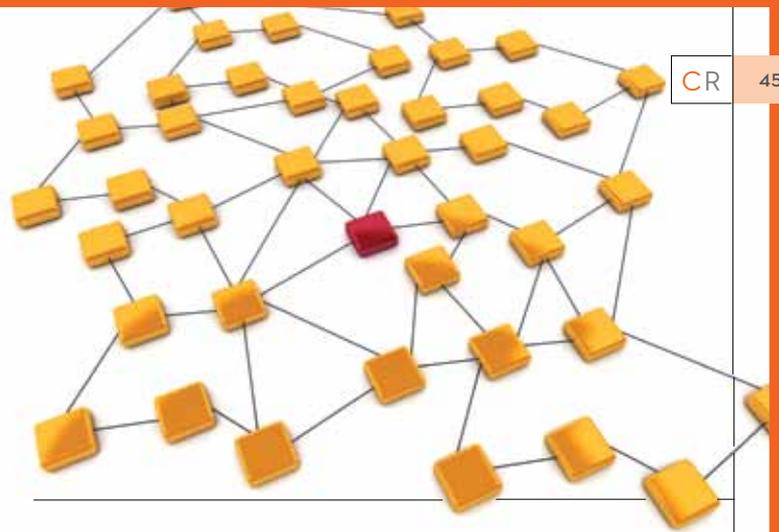
identities, it is far from comprehensive and is not open to the general public. The fact that humanities scholarship refers to work that can be decades and even centuries old was a theme of the bibliographic discussion at LAWDI.

One premise of LAWDI was that face-to-face interaction would lead to greater interlinking and reuse of digital resources in the future. LAWDI presenters frequently made the point that pointing to a stable digital resource is a form of endorsement that encourages yet more reuse of that same resource. This in turn can lead to interoperability of data as common identifiers form a basis for linking together disparate work. Essentially all LAWDI participants were eager to show resources that provide stable URLs or to ask for advice on what is currently available. But both the participants in and organizers of LAWDI recognize the need to take active steps to grow the number of high-quality digital resources. That will require ongoing outreach as well as clear examples of how Linked Open Data benefits both creators and users. As we plan for the 2013 session, it will be important to pay attention to tools that make it easier to take the first steps into publishing Linked Open Data. These tools may include cloud-based services such as GitHub.com, Nodester.com, and database hosting services such as Mongohq.com or Mongolabs.com. Likewise, technical developments such as RDFa and JSON-LD should increasingly take on their intended role as lower-cost entry points for Linked Data-based projects.

Basic information about LAWDI is available at the event's website, which is hosted by *The Digital Classicist*, a decentralized community that supports digital initiatives within Classical Studies. The page *LAWDI 2012 Websites* shows the very wide range of projects that participants are working on, some of which already implement Linked Open Data principles. The *LAWDI 2012 Documents Presentations* webpage has links to the extensive Twitter feed of the event, to a selection of the presenters' slides, and to blog posts that both preceded and followed LAWDI's three days of intensive conversation. These documents do much to capture the spirit of the event and to suggest ways that Linked Data will continue to transform research within the study of the Ancient World.

ICRI doi: 10.3789/isqv24n2-3.2012.08

**THOMAS ELLIOTT** (tom.elliott@nyu.edu) is Associate Director, Digital Programs and Senior Research Scholar, and **SEBASTIAN HEATH** (sebastian.heath@nyu.edu) is a Research Assistant Professor, both with the Institute for the Study of the Ancient World at New York University. **JOHN MUCCIGROSSO** (jmuccigr@gmail.com) is Associate Dean & Professor, Department of Classics, Drew University.



LAWDI presenters frequently made the point that pointing to a stable digital resource is a form of endorsement that encourages yet more reuse of that same resource. This in turn can lead to interoperability of data as common identifiers form a basis for linking together disparate work.

**Google Ancient Places**

[googleancientplaces.wordpress.com/](http://googleancientplaces.wordpress.com/)

**Library of Congress Linked Data Service**

[id.loc.gov/](http://id.loc.gov/)

**Linked Ancient World Data Institute**

[wiki.digitalclassicist.org/Linked\\_Ancient\\_World\\_Data\\_Institute](http://wiki.digitalclassicist.org/Linked_Ancient_World_Data_Institute)

**Open Annotation Collaboration**

[www.openannotation.org/](http://www.openannotation.org/)

**Pelagios**

[pelagios-project.blogspot.com/](http://pelagios-project.blogspot.com/)

**Perseus Digital Library**

[www.perseus.tufts.edu/](http://www.perseus.tufts.edu/)

**Pleiades**

[pleiades.stoa.org](http://pleiades.stoa.org)

**Virtual International Authority File (VIAF)**

[viaf.org/](http://viaf.org/)

**WorldCat**

[www.worldcat.org](http://www.worldcat.org)



**RELEVANT  
LINKS**



Kevin M.  
Ford

KEVIN M. FORD

## LC's Bibliographic Framework Initiative and the Attractiveness of Linked Data

With the Bibliographic Framework Initiative—a community effort led by the Library of Congress (LC) and first announced in 2011—the library world has begun its transition from the MARC 21 communication formats.

The MARC format is one of the oldest data format standards still in wide use today. Indeed, the format permeates everything in the library community: it is embedded in library technology and it is embedded in the minds of most librarians, especially catalogers, who know MARC and only MARC. It is undeniably part of the library family—it is the butt of jokes; it is the topic of conversations; it is worried about; it is cared for; it is loved; it is hated—and it is hard to envision life without MARC. It is, after all, forty-five years old. Most librarians working today began their careers after MARC was born, though they may have spent the first decade or two of their careers at a safe distance from the format. Some have never known life without MARC.

In 2011, LC started the initiative to phase out this library-technology stalwart and explore replacing it with a Linked Data model. The data model would, therefore, be grounded in Resource Description Framework (RDF), about which more is said below, and, in conjunction with an RDF model, the new framework would embrace the Linked Data practices and methods with respect to sharing and publishing library data. In this way, RDF provides a means to represent the data and the Linked Data methods and practices provide a means to communicate the data, the two core and historical functions of MARC.

### A brief history of MARC

The acronym stands for MACHine Readable Cataloging. The first tangible MARC project began at LC in January of 1966. The format—known as MARC I—was complete by April 1966, at which time testing to determine feasibility began. The fact

that the basic format was established in a four-month period is nothing short of astonishing. Around April 1967, work was underway revising the format, now called MARC II. It was formally published the following year. The LC MARC Distribution Service also began operation in 1968. In short, the MARC format was designed, tested, and implemented in little more than two years. LC led an aggressive development cycle that included a number of instrumental partners—Indiana University, the University of Chicago, Harvard University, and the National Agricultural Library, to name but a few—working in concert, testing and reporting their results back to LC. Incredibly, the MARC format, the second version, remained essentially unchanged for thirty years. It was in 1998 that the “21”—a nod to the rapidly approaching 21st century—was appended to “MARC,” marking the occasion when LC and the National Library of Canada merged their respective formats, the USMARC format and CAN/MARC. And so, today, we speak of MARC 21.

When working with MARC, one typically refers to a MARC record, which is a set of attributes and values that together independently describe a resource. Initially that resource was a book, but the MARC format was soon extended to describe other format types such as serials, maps, music, still images, and many more. In the mid-1970s these different format types went through some consolidation from which was born the more encompassing MARC Bibliographic Format. Following that, the MARC Authority format was formally published in 1981 (though LC had its authority data in an LC INTERNAL MARC format since about 1973); MARC Holdings format followed



Figure 1: A few simple RDF statements describing a book  
 ("ex" is an example namespace; "rdf" refers to the Resource Description Framework namespace; "dc" refers to Dublin Core)

a few years later. MARC Classification has existed since 1991. These various formats are collectively referred to as the MARC communication formats. Although structurally the formats adhere to the ISO 2709 standard (also standardized as ANSI/NISO Z39.2), official since 1973, each communication format employs its own codes and conventions to identify and describe the data elements within a record.

During the past forty-five year period, systems have been developed that permit catalogers to create bibliographic, authority, and classification records directly in the MARC format. In other cases, systems are at least capable of constructing and deconstructing MARC records even if the internal storage structure does not itself reflect the MARC 21 format. Programmers have written software to manipulate one or a group of MARC records, often by transforming the data from one format to another. Additional programs have been written that perform endless quality checks and other statistical analyses of MARC data. Libraries have all types of personnel who can look at and understand a MARC record as readily as they can read this sentence.

All of this is to say that the MARC format has had a long and productive life for the library community. By every measure, MARC has been a success, but it is antiquated when compared to our ability to model and store data in the second decade of the 21st century.

### The attractiveness of Linked Data

MARC was designed for the representation and communication of bibliographic and related information in

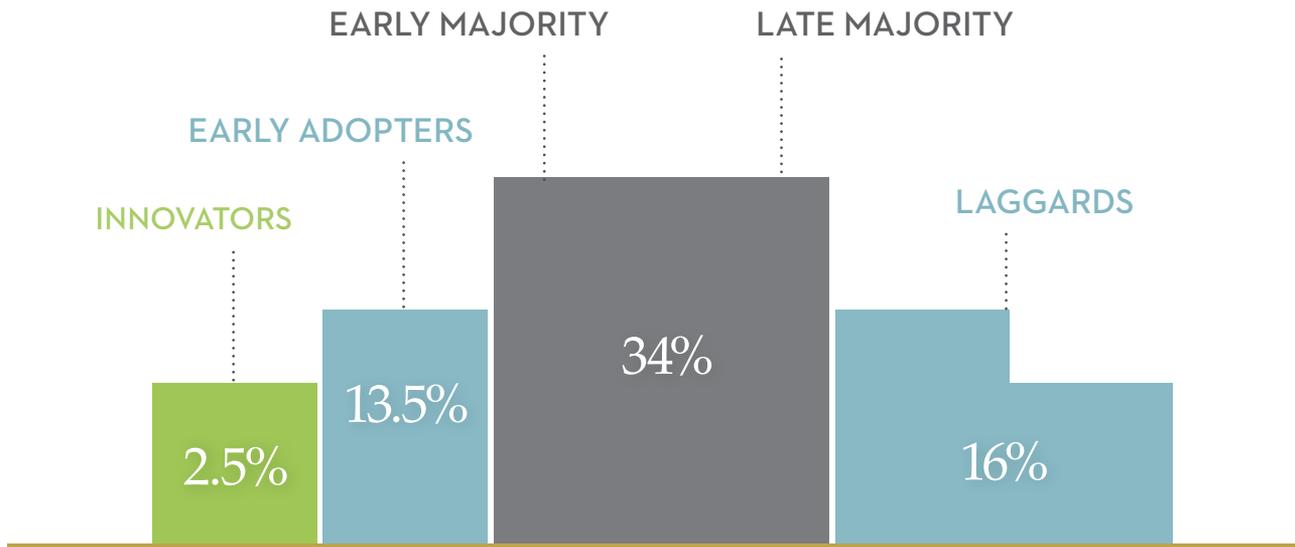
machine-readable form. Any replacement to MARC must be capable of performing those two functions: representation and communication. The knowledge, principles, practices, and technologies that have been developed for and that exist in support of Linked Data, and the accompanying movement, have made Linked Data a promising avenue of exploration. In its barest form, Linked Data is about publishing structured data over the same protocol used by the World Wide Web and linking that data to other data to enhance discoverability of more information.

Though not an absolute requirement, it is expected that the data conforming to Linked Data principles will be described using a very simple, but powerful, data model called the Resource Description Framework (RDF). Borrowing an analogy from English grammar, the parts of an RDF statement can be equated to those found in a basic linguistic sentence, which must contain a subject, verb, and, optionally, an object. In the case of RDF, the subject is a uniquely identified concept or thing (preferably with an HTTP URI, a uniform resource identifier), about which the statement is made. The other two parts are called the predicate (like a verb) and object. The predicate—also, identified with a URI—records the relationship between the subject and object. The object of the statement may be identified with a URI or it may be a string. It is possible to describe a thing or concept fully by making a number of RDF statements about it (see Figure 1.)

From Figure 1 we learn that the thing identified as "ex:Book12345" is a book by Giuseppe Moretti (died in 1945) about the Ara Pacis in Rome, Italy. Each of the lexical strings that are the objects of the above sentences could also be

CONTINUED »

FIGURE 2 : INNOVATION ADOPTION LIFECYCLE.



identifiers, which, when queried, would return a series of RDF statements where each identifier is the subject. Indeed, in a future bibliographic environment, the names and subjects, minimally, will be identifiers, thereby eliminating the current practice of embedding the lexical string into library records. If this alone were the only outcome of a new bibliographic environment (rest assured, there will be many more outcomes), then libraries and librarians would likely consider the enterprise a success.

RDF is a World Wide Web Consortium (W3C) recommendation (i.e., standard). The W3C is an international body that manages standards applicable to the Internet, and specifically the World Wide Web. HTML is the most well-known standard W3C has managed. RDF, which was formally published in 1999, comes under the umbrella of the W3C Semantic Web initiative. Although RDF itself is celebrating its thirteenth birthday, it has only been in the last five years that storage software, software tools, and general knowledge about RDF have matured beyond the “innovation” stage and penetrated safely into the relatively more confident “early adopters” stage, per the technology adoption lifecycle developed by Bohlen, Beal, and Rogers, shown in Figure 2.

Libraries, librarians, and developers have been active innovators throughout the thirteen year period of RDF’s ascendancy, during which time much has been tried and

much has been learned—all of which directly informs current thinking about the Bibliographic Framework Initiative. One of the first instances of the Library of Congress publishing RDF came in 2005 when the MARC Relators list (a list of codes used to identify the role an individual had in the lifecycle of a resource, such as a book) was mapped to Dublin Core properties and published online. Although not the first of its kind, the LC Linked Data Service—[id.loc.gov](http://id.loc.gov)—went online in early 2009 and featured the LC Subject Headings file in RDF.

Libraries have legacy systems, legacy practices, and legacy data that must be carefully managed through any and all transitions. As mentioned in earlier announcements and articles about the Bibliographic Framework Initiative, the transition away from MARC will not be revolutionary, but a gradual process that ensures data integrity, system stability, and that no group is unintentionally left behind, in so far as is manageable. That RDF technologies may be on the cusp of entering the early majority stage means the library community is moving at the right time. It has only been in the last five years, or so, that RDF technologies have matured, including the uptake of those technologies, such that there is sufficient confidence in their continued support, development, and improvement. These technologies range from RDF vocabularies to RDF software libraries and to robust and scalable RDF triplestores (databases

specifically designed to store and query RDF data). The library community needs to migrate to a new framework in a conscientious and responsible manner that accounts for the current state of library technology, but we also desire to enjoy a technological future that does not require costly, complete redevelopment every decade.

The fact that RDF is published by and managed by a standards organization like the W3C means that many developers and technologists beyond the library sector will more easily understand library data formats and technology. This is not the case today. Any RDF vocabulary or ontology developed in support of library data will, of course, still require subject matter expertise to fully understand its semantics, but developers and programmers will not need to understand the structural design of an ISO 2709 MARC record. Moreover, because the Bibliographic Framework Initiative is grounded in well-known and well-understood standards and technology that are widely used beyond the library sector, more individuals and companies will be competing in this space. Libraries will have a greater selection of services and solutions from which to choose.

Beyond the technology surrounding and supporting RDF, Linked Data methods and principles coincide perfectly with the mores and practices of libraries. Linked Data is about sharing data (i.e., publishing data). Developers have identified, promoted, and coalesced around an entire set of technical procedures—all grounded in the HTTP protocol—that have become commonly accepted (and expected) practice to facilitate access to structured RDF data via the World Wide Web. Dereferenceable URIs and content negotiation are two such procedures (neither belonging exclusively to the domain

of Linked Data). Not only do these methods help to expose library data widely and make it more accessible, but also the Linked Data movement provides a strong and well-defined means to *communicate* library data, one of the main functions requiring attention in the community's migration from MARC. Perhaps most importantly, by pursuing the Linked Data model for information sharing, the future Bibliographic Framework will embrace the notion of "The Network."

Instead of library information having to be specially extracted from backend databases, packaged as groups of records (or singly), and then made intentionally available via some kind of transfer protocol, The Network will be the center of the model. Today, library records are independently understandable; an author's name is in the record, as are the subjects, for example. In the future, there may be only an opaque identifier—a reference, a link, an HTTP URI—

CONTINUED »

---

Transition away from MARC will not be revolutionary, but a gradual process that ensures data integrity, system stability, and that no group is unintentionally left behind, in so far as is manageable.



to a resource that further describes the author or subject. It will be by performing content negotiation on the dereferenceable HTTP URI that a requesting agent (a human or system) will learn the lexical, human-readable string that is the actual name or subject heading associated with the identifier. More likely, systems will maintain a local copy of authority data, such as names and subjects, not only for indexing purposes but also because this is more efficient than requesting the lexical value over the network millions of times a day. But the maintenance, and the sharing, of this type of information will be fluid and infinitely easier in a Linked Data model than it is today.

### Where to now?

The library community will need to further refine, customize, and possibly standardize (at least for faithful operation within the library community) the technology methods surrounding the exchange of and (potentially) the representation of library data in order to fully realize the Linked Data approach, with accompanying RDF model, in a new Bibliographic Framework. Work to date has revealed that technically conformant linked data service installations such as LC's Linked Data Service will require expansion and refinement to serve the greater requirements of the new Bibliographic Framework. Although the present Linked Data methods are technically satisfactory, additional services can be implemented to ameliorate server loads, client processing energy, and a host of other small issues that singly amount to little but in the aggregate quickly become issues of scale. For example, a simple URI-to-string service would permit a client to request the lexical, human-readable value for a URI without the programmatic drudgery of sifting through every statement about a resource (especially when only one is needed), which is the current result based on Linked Data practices.

As characterized above, this is a transition from one bibliographic framework to a new one. Our legacy systems today deserve careful consideration and a measured approach. The timeline pursued by LC and its partners from January 1966 to the creation of the MARC Distribution Service in 1968 will not, therefore, be nearly as aggressive. Nevertheless, work has begun on developing a model for community discussion and identifying the technology needs to support the model. In the end, however, LC will still require—and, more importantly, wants—partners for this effort. There will be much for the community to contribute. LC is taking the lead with the Bibliographic Framework Initiative by coordinating and managing it, but the Initiative's success rests on the valuable contributions (already received and still to come) from the wider community in the forms of discussion, feedback, testing, and, above all, participation during this process. | SP | doi: 10.3789/isqv24n2-3.2012.09

**KEVIN M. FORD** ([kefo@loc.gov](mailto:kefo@loc.gov)) works in the Network Development and MARC Standards Office, Library of Congress, and is the project manager for the LC Linked Open Data service.



Work has begun on developing a model for community discussion and identifying the technology needs to support the model. In the end, however, LC will still require—and, more importantly, wants—partners for this effort. There will be much for the community to contribute.

**LC Bibliographic Framework  
Transition Initiative**  
[www.loc.gov/marc/transition/](http://www.loc.gov/marc/transition/)

**LC Linked Data Service**  
[id.loc.gov/](http://id.loc.gov/)

**MARC Standards**  
[www.loc.gov/marc/](http://www.loc.gov/marc/)

**Resource Description Framework  
(RDF) Primer**  
[www.w3.org/TR/2004/REC-rdf-primer-20040210/](http://www.w3.org/TR/2004/REC-rdf-primer-20040210/)



**RELEVANT  
LINKS**



Nettie  
Lagace

## Updated Recommended Practice on SERU: A Shared Electronic Resource Understanding

A new edition of the recommended practice *SERU: A Shared Electronic Resource Understanding* (NISO RP-7-2012) expands use of SERU beyond e-journals. The SERU Recommended Practice offers a mechanism that can be used as an alternative to a license agreement by expressing commonly shared understandings between content providers and libraries. These understandings include such things as the definition of authorized users, expectations for privacy and confidentiality, and online performance and service provisions. The 2012 updated version of SERU recognizes both the importance of making SERU more flexible for those who want to expand its use beyond e-journals, while acknowledging the fact that consensus for other types of e-resource transactions are not as well established as they are for e-journals.

Since the 2008 publication of the original SERU RP, many models have emerged for acquiring e-books and both libraries

and e-book providers have requested that other types of electronic resources be incorporated into the SERU framework. This new version uses language that can be applied to a wide variety of e-resources while retaining the same shared understandings that made the previous version so useful.

The SERU Registry of those interested in using the SERU approach already contains over 70 publishers and content providers and 185 libraries and consortia. The expansion of the recommendations to address additional types of e-resources should interest more organizations in joining the SERU registry.

 The SERU Recommended Practice, the SERU Registry, and additional helpful resources are available from the SERU workroom webpage on the NISO website: [www.niso.org/workrooms/seru/](http://www.niso.org/workrooms/seru/).

## New Authoring and Interchange Framework Standard

NISO and the DAISY Consortium announced the publication in August 2012 of the new American National Standard *Authoring and Interchange Framework* (ANSI/NISO Z39.98-2012). The standard defines how to represent digital information using XML to produce documents suitable for transformation into different universally accessible formats. The standard is a revision, extension, and enhancement of *Specifications for the Digital Talking Book (DTB)*, commonly referred to as the DAISY standard (ANSI/NISO Z39.86-2005 (R2012)). The DAISY Consortium is the Maintenance Agency for both standards.

The A&I Framework is a modular, extensible architecture to permit the creation of any number of content representation models, each custom-tailored for a particular kind of information resource. It also provides support for new output formats, which can be added and implemented as the need arises. The standard does not impose limitations on what distribution formats can be created from it; e-text, Braille, large print, and EPUB are among formats that can be produced in conformance with the standard.

Although the new A&I Framework standard is intended to replace the Digital Talking Book standard, feedback during trial use of the standard indicated that content providers and device manufacturers would need a transition period of several years due to the significance of the changes in the standard. To meet this need, the existing DTB standard (ANSI/NISO Z39.86) was reaffirmed for another five years and the A&I Framework was assigned a new standard number (ANSI/NISO Z39.98).

The A&I Framework standard will be of interest to any organization using an XML authoring workflow, developers and publishers of universally accessible digital publications, and agencies interested in creating profiles for new document types to integrate into distribution formats, such as EPUB.

 Both the A&I Framework standard and the Digital Talking Book standard are available for free download from the NISO website ([daisy.niso.org](http://daisy.niso.org)) and the DAISY website ([www.daisy.org/daisy-standard](http://www.daisy.org/daisy-standard)).

## New Initiative to Develop Recommended Practices for Demand-Driven Acquisition (DDA) of Monographs

NISO voting members approved a new project to develop recommended practices for the Demand-Driven Acquisition (DDA) of Monographs. Many libraries have embraced DDA (also referred to as patron-driven acquisition) to present many more titles to their patrons for potential use and purchase than would ever be feasible under the traditional purchase model. If implemented correctly, DDA can make it possible to purchase only what is needed, allowing libraries to spend the same amount of money as they previously spent on monographs, but with a higher rate of use. However, this model requires libraries to develop and implement new procedures for adding titles to a “consideration pool,” for keeping un-owned titles available for purchase for some future period (often years after publication), for providing discovery methods of titles in the pool, establishing rules on when a title gets purchased or only

temporarily leased, how potential titles are discovered, and for handling of multiple formats of a title.

DDA can be a significant disruption in the existing supply chain for monographs, not only for libraries but also for publishers, sales agents, aggregators, and end users. New roles and practices need to be shaped in a way that allows the scholarly communication supply chain to continue to function effectively. Additionally, most libraries that have experimented with DDA have been in the academic sector; NISO intends to involve the public library community with this project and develop recommendations that can work for all library types.

 More information about the project, including the project proposal can be found on the NISO DDA workroom webpage: [www.niso.org/workrooms/dda/](http://www.niso.org/workrooms/dda/)

## Process Begun for National Standardization of the 3M Standard Interchange Protocol (SIP)

NISO voting members have approved a new project to formalize the 3M Standard Interchange Protocol (SIP) as an American National Standard. Introduced in 1993, the SIP protocol provides a mechanism for Integrated Library Systems (ILS) applications and self-service devices to communicate seamlessly to perform self-service transactions. This protocol quickly became a de facto standard around the world, and remains the primary protocol to integrate ILS and self-service devices. Since the protocol's inception, 3M has continued to produce updated versions of it—most recently version 3.0 in late 2011. A NISO Working Group will now shepherd SIP 3.0 through the standardization process of becoming an American National Standard.

“While 3M has always sought input from the library community of developers and interested parties in enhancing the protocol, the time is right for further maintenance and upgrades to SIP to be done in a more independent, community environment,” stated Sue Boettcher, 3M Senior Product Development Specialist. “3M will continue to participate, but as a contributing vendor and user of the protocol.”

“Obviously, there is close connection between SIP and NISO’s Circulation Interchange Protocol (NCIP) standard,” said Robert Walsh, representative for EnvisionWare, the Maintenance Agency for NCIP. “With both standards approved and maintained within NISO, there is an opportunity for the two standards’ working groups to clarify the structural differences and to provide the community direction on the appropriateness for each standard within a given context. This will be one of the tasks of both the new working group and the NCIP Maintenance Agency moving forward.”

 More information about the project, including the project proposal can be found on the NISO SIP workroom webpage: [www.niso.org/workrooms/sip](http://www.niso.org/workrooms/sip).



I NR | doi: 10.3789/isqv24n2.2012.10



The Draft Release 1 of the COUNTER Code of Practice for Usage Factors, is one of the most significant outcomes to-date of the Usage Factor project, and is an important part of the final stage of the project, which will take Usage Factor forward to full implementation.

## Draft Release 1 of the COUNTER Code of Practice for Usage Factors

The overall aim of the Usage Factor project is to explore how online journal usage statistics might form the basis of a new measure of journal impact and quality, the Usage Factor (UF). The Draft Release 1 of the COUNTER Code of Practice for Usage Factors, is one of the most significant outcomes to-date of the Usage Factor project, and is an important part of the final stage of the project, which will take Usage Factor forward to full implementation.

COUNTER's purpose in publishing this Draft Release 1 now, is threefold:

- 1 First, it sets out a formal, detailed standard for the recording, reporting, and maintenance of Usage Factors, solidly based on the outcomes of Stages 1 and 2 of the Usage Factor project
- 2 Second, it provides a document for interested parties to review and comment upon, which we encourage, as this will greatly help us develop a definitive, implementable Code of Practice.
- 3 Third, it provides a framework for selected publishers and other organizations to do more extensive testing of the proposed processes for recording and reporting Usage Factors

The draft will be available for comment on the COUNTER website until September 30, 2012. Comments should be sent to Peter Shepherd, COUNTER Director. ■

🔗 COUNTER COP for Usage Factors: [www.projectcounter.org/usage\\_factor.html](http://www.projectcounter.org/usage_factor.html)

## Ringgold and Bowker Appointed as ISNI Registration Agencies for Institutions

The ISNI International Agency (ISNI-IA) that is responsible for the administration of ISO 27729, *International Standard Name Identifier (ISNI)*, has appointed Ringgold, Inc. as the first registration agency for assignment of ISNIs to institutions. Ringgold will incorporate ISNIs into its *Identify* database of institutional identifiers and will provide a free look-up service at [www.openidentify.com](http://www.openidentify.com) which, after registration, enables users to search for and obtain an institutional identification number. The ISNI-IA has also appointed Bowker as the first U.S. registration agency for ISNIs. Bowker, an affiliated business of ProQuest, will be assigning ISNIs across the standard's scope in addition to institutions.

The ISNI standard was developed as an identifier for parties involved throughout the media content industries, including authors, musicians, publishers, rights holders, and even fictional characters. NISO's Institutional Identifier (I<sup>2</sup>) Working Group reached agreement with the ISNI

International Agency to apply the ISNI to institutions in addition to the originally attended scope, rather than developing an additional identifier standard just for institutions.

The ISNI International Agency (ISNI-IA) was founded by CISAC, the Conference of European National Librarians (represented by the Bibliothèque Nationale de France and the British Library), IFRRO, IPDA, OCLC, and ProQuest and appointed by ISO to administer implementation of the ISNI standard. ■

Ⓢ ISNI International Agency: [www.isni.org](http://www.isni.org)

NISO I<sup>2</sup> Working Group: [www.niso.org/workrooms/i2](http://www.niso.org/workrooms/i2)

Ringgold: [www.ringgold.com/](http://www.ringgold.com/)

Bowker: [www.bowker.com/](http://www.bowker.com/)



## Unified Digital Format Registry Launched

The University of California Curation Center (UC3) at the California Digital Library (CDL) announced the availability of the Unified Digital Format Registry (UDFR), a new semantically-enabled, community-supported open source platform for the collection, long-term management, and dissemination of the significant properties of formats of interest to the preservation community. A deep understanding of digital formats is necessary to support the long-term preservation of digital assets, as it facilitates the preservation of the information content of those assets, rather than just their bit stream representations. A format is the set of syntactic and semantic rules that govern the mapping between information and the bits that represent that information. The UDFR is expected to become a key piece of preservation infrastructure of use to the international preservation, curation, and repository communities.

The UDFR builds upon and “unifies” the function and holdings of two existing registry solutions: PRONOM, from the UK National Archives; and GDFR (Global Digital Format Registry), from Harvard University. While these services rely on older relational and XML database technology, the UDFR uses a semantic database in which all information is represented in RDF form and exposed as Linked Data

for interoperability with the evolving semantic web. Use of the UDFR is open to the public, although contribution or editing of information requires prior self-service account registration.

The UDFR was developed by UC3 with funding from the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP). ■

Ⓢ The UDFR is available at: [udfr.org/](http://udfr.org/)

---

**Use of the UDFR is open to the public, although contribution or editing of information requires prior self-service account registration.**

---





## W3C Launches Linked Data Platform Working Group

The World Wide Web Consortium announced the formation in May 2012 of a new Linked Data Platform (LDP) Working Group to promote the use of linked data on the Web. Per its charter, the group will explain how to use a core set of services and technologies to build powerful applications capable of integrating public data, secured enterprise data, and personal data. The platform will be based on proven Web technologies including HTTP for transport, and RDF and other Semantic Web standards for data integration and reuse. The group will produce supporting materials, such as a description of uses cases, a list of requirements, and a test suite and/or validation tools to help ensure interoperability and correct implementation.

The group's work is intended to complement SPARQL and bring the data integration features of RDF to RESTful, data-oriented software development. One or more W3C Recommendations will be produced that define a RESTful way to read and write Linked Data, suitable for use in application integration and the construction of interoperable and modular software systems. First public Working Drafts are expected to be published in October 2012. ■

 LDP Working Group Charter:  
[www.w3.org/2012/ldp/charter](http://www.w3.org/2012/ldp/charter)

## MARBI Disbanded; New ALCTS/LITA Metadata Standards Committee Formed

The Library and Information Technology Association and the Association for Library Collections & Technical Services (ALCTS), with the support of Reference and User Services Association (RUSA)—all divisions of the American Library Association—have formed the ALCTS/LITA Metadata Standards Committee, according to Zoe Stewart-Marshall, LITA 2012/13 President.

Marshall states that, “the ALCTS/LITA Metadata Standards Committee will play a leadership role in the creation and development of metadata standards for bibliographic information. The Committee will review and evaluate proposed standards; recommend approval of standards in conformity with ALA policy; establish a mechanism for the continuing review of standards (including the monitoring of further development); provide commentary on the content of various implementations of standards to concerned agencies; and maintain liaison with concerned units within ALA and relevant outside agencies.”

In announcing the formation of the new standards committee, Marshall also said that “the three ALA divisions have also voted to disband the ALCTS/LITA/RUSA Machine-Readable Bibliographic Information (MARBI) Committee, as of June 30, 2013. After June 2013, the MARC Advisory Committee (MAC) is expected to continue to advise the Library of Congress on MARC development. While there will no longer be MARBI involvement with MAC, other ALA representatives and liaisons as noted on the MAC roster will continue to advise LC about MARC. If a major issue related to MARC requires the attention of a voting ALA body, the issue may be brought to the new ALCTS/LITA Metadata Standards Committee. MARC, however, is not expected to be the prevailing focus of the new ALCTS/LITA committee. For the past several decades, MARBI has played a critical role in improving library metadata, particularly the MARC formats. ALCTS, LITA, and RUSA thank all those who have contributed to MARBI's many accomplishments. We look forward to working with the metadata community broadly in developing and monitoring current and emerging metadata standards.”

The Metadata Standards Committee will begin its work at the Midwinter Meeting of the American Library Association, January 2013. ■

 LITA: [www.ala.org/lita/](http://www.ala.org/lita/)

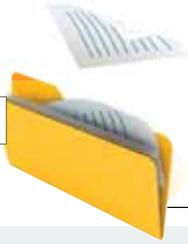
ALCTS: [www.ala.org/alcts/](http://www.ala.org/alcts/)

MARC Advisory Committee: [www.loc.gov/marc/marbi/advisory2.html](http://www.loc.gov/marc/marbi/advisory2.html)

I NW I doi: 10.3789/isqv24n2-3.2012.11



STAY UP-TO-DATE ON  
NISO NEWS & EVENTS:  
[www.niso.org/news](http://www.niso.org/news)



Listed below are the NISO working groups that are currently developing new or revised standards, recommended practices, or reports. Refer to the NISO website ([www.niso.org/workrooms/](http://www.niso.org/workrooms/)) and the Newsline quarterly supplements, *Working Group Connection* ([www.niso.org/publications/newsline/](http://www.niso.org/publications/newsline/)), for updates on the working group activities.

Note: DSFTU stands for Draft Standard for Trial Use.

WORKING GROUP	STATUS
<b>AI Framework (was DAISY Revision)</b> Co-chairs: Markus Gylling, George Kerscher	ANSI/NISO Z39.98-2012, <b>Authoring and Interchange Framework for Adaptive XML Publishing Specification</b> Approved for publication 7/9/2012.
<b>Demand Driven Acquisition of Monographs</b> Co-chairs: TBA	Working Group being formed.
<b>Digital Bookmarking and Annotation Sharing</b> Co-chairs: Baden Hughes, Dan Whaley	NISO Z39.97-201x, <b>Digital Bookmarking and Annotation Standard</b> in development.
<b>E-book Special Interest Group</b> Co-chairs: Nettie Lagace, Todd Carpenter	Pre-standardization work underway in four sub-groups: Accessibility, Discovery Tools and Linking, Distribution, Metadata
<b>Institutional Identifiers (I<sup>2</sup>)</b> Co-chairs: Grace Agnew, Oliver Pesch	NISO RP-17-201x, <b>Institutional Identification: Identifying Organizations in the Information Supply Chain.</b> Finalizing for publication.
<b>Improving OpenURLs Through Analytics (IOTA)</b> Chair: Adam Chandler	Technical Report in development.
<b>Knowledge Base and Related Tools (KBART) Phase II</b> <i>Joint project with UKSG.</i> Co-chairs: Andreas Biedenbach, Sarah Pearson	Phase II Recommended Practice in development.
<b>Open Discovery Initiative</b> Co-chairs: Marshall Breeding, Jenny Walker	Recommended Practice in development.
<b>Presentation and Identification of E-Journals (PIE-J).</b> Co-chairs: Bob Boissy, Cindy Hepfer	NISO RP-16-201x, <b>PIE-J: The Presentation &amp; Identification of E-Journals</b> Finalizing for publication following the public comment period.
<b>Resource Synchronization</b> Co-chairs: Herbert Van de Sompel, Todd Carpenter	NISO Z39.99-201x, <b>Specification for Web Resource Synchronization</b> In development.
<b>Standard Interchange Protocol (SIP)</b> Co-chairs: John Bodfish, Ted Koppel	NISO Z39.100-201x <b>Standard Interchange Protocol (SIP)</b> Working group being formed.
<b>Standardized Markup for Journal Articles</b> Co-chairs: Jeff Beck, B. Tommie Usdin	NISO Z39.96-2012, <b>JATS: Journal Article Tag Suite, version 1.0</b> Approved by NISO; ANSI approval pending.
<b>Supplemental Journal Article Materials</b> <i>Joint project with NFAIS.</i> Co-chairs Business Working Group: Linda Beebe, Marie McVeigh. Co-chairs Technical Working Group: Dave Martinsen, Alexander (Sasha) Schwarzman	NISO RP-15-201x, <b>Recommended Practices for Online Supplemental Journal Article Materials</b> Part A: Business Working Group Recommendations being finalized for publication following draft for comments. Part B: Technical Working Group Recommendations issued as a Draft for Comments ending September 15, 2012.
<b>SUSHI Server Working Group.</b> Chair: Oliver Pesch	NISO RP-13-201x, <b>Providing a Test Mode for SUSHI Servers</b> Finalizing for publication following a draft for trial use.
<b>SUSHI (Z39.93) Standing Committee</b> Co-chairs: Bob McQuillan, Oliver Pesch	NISO RP-14-201x, <b>NISO SUSHI Protocol: COUNTER-SUSHI Implementation Profile</b> Finalizing for publication following a public comment period.
<b>Z39.83 (NCIP) Standing Committee</b> Co-chairs: Mike Dicus, Robert Walsh	NISO Z39.83-1-2012, <b>NISO Circulation Interchange Part 1: Protocol (NCIP), version 2.02</b> NISO Z39.83-2-2012, <b>NISO Circulation Interchange Protocol (NCIP) Part 2: Implementation Profile 1</b> Approved by NISO; ANSI approval pending.

# TRACKING IT BACK TO THE SOURCE: MANAGING AND CITING RESEARCH DATA

SEPTEMBER 24, 2012 | DENVER, CO

As data creation increases exponentially across nearly all scholarly disciplines, new roles and requirements are rising to meet the challenges in organization, identification, description, publication, discovery, citation, preservation, and curation to allow these materials to realize their potential in support of data-driven, often interdisciplinary research.

This Forum will focus on several new initiatives to improve community practice on data citation and data discovery.

 For information and to register visit:  
[www.niso.org/news/events/2012/tracking\\_it\\_back\\_to\\_the\\_source/](http://www.niso.org/news/events/2012/tracking_it_back_to_the_source/)



## LEARN ABOUT:

- » DataCite and EZID
- » Data Equivalence
- » ResourceSync: the Large-Scale Synchronization of Web Resources
- » Data Observation Network for Earth (DataONE)
- » Data Attribution and Citation Practices





**NISO  
2-DAY  
Forum**

# THE E-BOOK RENAISSANCE PART II: CHALLENGES AND OPPORTUNITIES

OCTOBER 18-19, 2012

---

METRO MEETING CENTERS | BOSTON, MA

---

Join us at the NISO Forum on *E-book Renaissance Part II: Challenges and Opportunities* as we explore availability, access, distribution, licensing, discoverability, and usage issues from a variety of industry, scholarly, and consumer viewpoints. Participate in the community discussion to advance e-book development and support.

E-books have existed in the library landscape for over a decade, but it is only in the last few years that their use has grown to finally become the game-changer that many have anticipated for so long. E-book availability, distribution, licensing, discoverability, usage, and current and future access require content providers and libraries to change many of their existing processes and develop new ways to do business. Amidst this confusion is a wealth of opportunities for new collaborations and initiatives.



**FOR INFORMATION AND TO REGISTER, VISIT:**

[www.niso.org/news/events/2012/ebooks](http://www.niso.org/news/events/2012/ebooks)