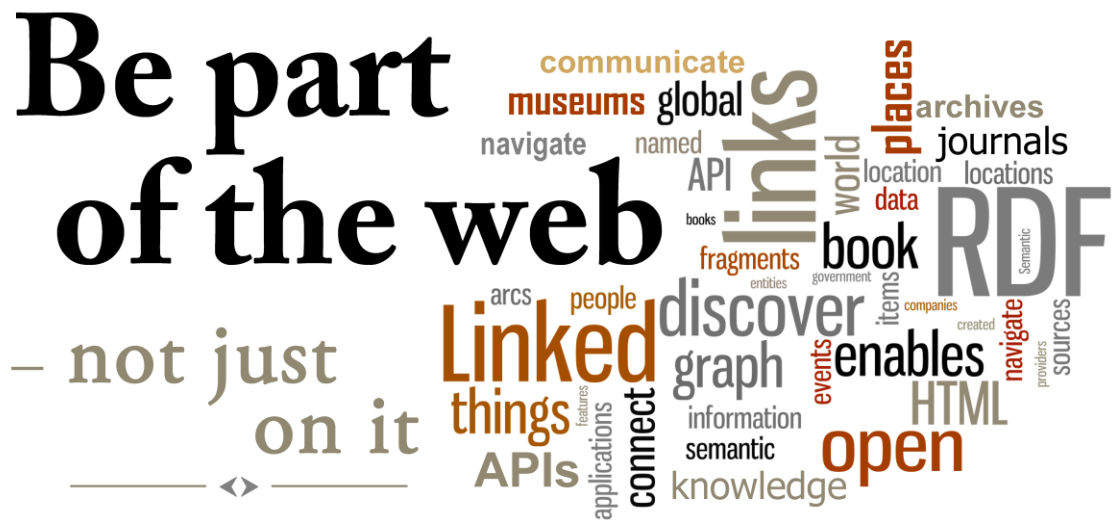


Report of the Stanford Linked Data Workshop, 27 June – 1 July 2011

ABSTRACT

The Stanford University Libraries and Academic Information Resources (SULAIR) with the Council on Library and Information Resources (CLIR) conducted at week-long workshop on the prospects for a large scale, multi-national, multi-institutional prototype of a Linked Data environment for discovery of and navigation among the rapidly, chaotically expanding array of academic information resources. As preparation for the workshop, CLIR sponsored a survey by Jerry Persons, Chief Information Architect emeritus of SULAIR that was published originally for workshop participants as background to the workshop and is now publicly available. The original intention of the workshop was to devise a plan for such a prototype. However, such was the diversity of knowledge, experience, and views of the potential of Linked Data approaches that the workshop participants turned to two more fundamental goals: building common understanding and enthusiasm on the one hand and identifying opportunities and challenges to be confronted in the preparation of the intended prototype and its operation on the other. In pursuit of those objectives, the workshop participants produced:

1. a value statement addressing the question of why a Linked Data approach is worth prototyping;
2. a manifesto for Linked Libraries (and Museums and Archives and …);

3. an outline of the phases in a life cycle of Linked Data approaches;
4. a prioritized list of known issues in generating, harvesting & using Linked Data;
5. a workflow with notes for converting library bibliographic records and other academic metadata to URIs;
6. examples of potential "killer apps" using Linked Data: and
7. a list of next steps and potential projects.

This report includes a summary of the workshop agenda, a chart showing the use of Linked Data in cultural heritage venues, and short biographies and statements from each of the participants.

This report was compiled by Michael A. Keller, Jerry Persons, Hugh Glaser, and Mimi Calter. It was published October 2011.

The accompanying survey is available at http://www.clir.org/pubs/archives/linked-data-survey/ .

CONTENTS

INTRODUCTION

From 27 June to 1 July 2011, Stanford University hosted a group of librarians and technologists to confront the challenge of planning a multi-national, multi-institutional discovery environment based on the use of Linked Data. It was foreseen that part of the workshop would involve the identification and examination of the issues and stumbling blocks around the use of Linked Data for academic library applications. All participants had some involvement in either the Linked Data arena or library metadata, though their backgrounds and experiences differ dramatically. Nevertheless, the participants shared a vision of Linked Data as disrupter technology with the potential to move libraries and other information providers beyond the restrictions of MARC based metadata as well as the restrictions of many variant forms of metadata generated for the wide variety of genres in use in scholarly communication. The participants in the workshop endorsed the need to precipitate a new family of tools and services featuring an array of emergent, open, link-driven meta-services in order to enable fully Linked Data as a disrupter technology for discovery, navigation, and business processes.

The stated objective of the workshop was the creation of fundable plans for the development of such tools and the definition of a prototype environment that would demonstrate the viability of the Linked Data approach. In the early stages of the workshop itself, however, it became clear that the identification and explication of use cases for such tools, as well as the identification of key stumbling blocks for their implementation, were objectives that by necessity took precedence over creating a plan. In addition, the Workshop sought to identify partners, either among the workshop participants or beyond them, to take on various aspects of projects identified within the workshop.

This report

- details the products of the workshop;
- outlines the next steps identified by the participants including achieving the objective of creating a proposal for a Linked Data prototype environment;
- provides biographies of the workshop participants; and
- summarizes the activities and discussions that took place during the workshop.

Two additional objectives arising from the workshop discussion are these:

1) The workshop participants should identify projects that we could reasonably accomplish from the resources over which we have direct influence. These would be "lighthouse projects" that would exemplify different aspects of what could be achieved and ideally at least some of those would be of sufficient size and cover numerous genres to demonstrate the vitality of Linked Data environments for discovery and navigation for information objects whose metadata and even full texts are contained in numerous separate silos. See pp. 45-48 of this report.

2) The group as a whole should find a way to encourage linked date activity from the broader community. This could be done by means of tutorials, references to technologies and methodologies and a framework in which the larger community could contribute. Being able to lower the barrier to entry (to the Linked Data world) for institutions that hold unique data seems to be a key success factor. The Value Statement (p. 15) and the Manifesto (p. 17) are elements of this advocacy.

Themes that ran through the discussion include the need to move beyond proprietary tools, services, and environments, and develop tools that truly would be truly open and unencumbered by proprietary interests. [1]

The Stanford Linked Data Workshop was co-sponsored by the Council on Library and Information Resources (CLIR) and the Stanford University Libraries and Academic Information Resources (SULAIR) with funding from the Andrew W. Mellon Foundation, CLIR, and SULAIR.

[1] There was a running tension at the workshop and in the development of this document between the need to "throw-up the URI as soon as possible" (manifesto no. 2) and the need for accuracy and curation (note on high confidence after Workflows 6). Actually, publishing one's own URIs is a recipe for the most accuracy – you say what you want about exactly what you want. Institutions and individuals should not to be afraid of minting new URIs, and certainly not delay the process of "Triplification " trying to do it. Where a publisher, meaning a minter of URIs, has strong and stable identifiers (URIs) of their own already, then these should be used. Otherwise, it is worth putting in the effort to find if there are other strong and stable identifying URIs to which resources can be **easily and reliably** mapped.

QUOTE FROM LIBRARY LINKED DATA INCUBATOR GROUP FINAL REPORT

W3C INCUBATOR GROUP REPORT[2]

[Compiler's note: The following quote provides a strategic introduction from the WC3 draft report of relevance to leaders of libraries and their advisors. Please note that the quote focuses upon metadata produced by libraries, which was only one of the many foci of the Stanford Linked Data Workshop. In principle, the "Benefits" and "Current Situations" sections of the following quote are apropos to libraries, archives, and museums. The quote begins here and concludes on p. 14.]

Scope of this Report

The scope of this report -- "library Linked Data" -- can be understood as follows:

Library. The word "library" as used in this report comprises the full range of cultural heritage and memory institutions including libraries, museums, and archives. The term refers to three distinct but related concepts: a collection of physical or abstract (potentially including "digital") objects, a place where the collection is located, and an agent that curates the collection and administers the location. Collections may be public or private, large or small, and are not limited to any particular types of resources.

Library data. "Library data" refers to any type of digital information produced or curated by libraries that describes resources or aids their discovery. Data covered by library privacy policies is generally out of scope. This report pragmatically distinguishes three types of library data based on their typical use: **datasets**, **element sets**, and **value vocabularies** (see Appendix A)

Linked Data. "Linked Data" refers to data published in accordance with principles designed to facilitate linkages among datasets, element sets, and value vocabularies. Linked Data uses Uniform Resource Identifiers (URIs) as globally unique[3] identifiers for

---

[2] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

[3] One deeply involved participant at the Workshop observes: "URIs are not "unique", or at least it is deeply open to misunderstanding to describe them as such. They are unambiguous, in the sense they relate to a single resource,but the idea of a "unique identifier" might well be understood to mean a resource only has one identifier – this is the reverse mapping. In some sense, any identifier is unique, just as any one thing is unique.

any kind of resource -- analogously to how identifiers are used for authority control in traditional librarianship. In Linked Data, URIs may be Internationalized Resource Identifiers (IRIs) -- Web addresses that use the extended set of natural-language scripts supported by Unicode. Linked Data is expressed using standards such as Resource Description Framework (RDF), which specifies relationships between things -- relationships that can be used for navigating between, or integrating, information from multiple sources.

Open Data. While "Linked Data" refers to the technical interoperability of data, "Open Data" focuses on its legal interoperability. According to the definition for Open Bibliographic Data, Open Data is in essence freely usable, reusable, and redistributable -- subject, at most, to the requirements to attribute and share alike. Note that Linked Data technology per se does not require data to be Open, though the potential of the technology is best realized when data is published as Linked Open Data.

Library Linked Data. "Library Linked Data" is any type of library data (as defined above) that is expressed as Linked Data.

Benefits

Benefits of the Linked Data Approach

The Linked Data approach offers significant advantages over current practices for creating and delivering library data while providing a natural extension to the collaborative sharing models historically employed by libraries. Linked Data and especially Linked Open Data is **sharable, extensible,** and easily **re-usable.** It supports multilingual functionality for data and user services, such as the labeling of concepts identified by a language-agnostic URIs. These characteristics are inherent in the Linked Data standards and are supported by the use of Web-friendly identifiers for data and concepts. Resources can be described in collaboration with other libraries and linked to data contributed by other communities or even by individuals. Like the linking that takes place today between Web documents, Linked Data allows anyone to contribute unique expertise in a form that can be reused and recombined with the expertise of others. The use of identifiers allows diverse descriptions to refer to the same thing.

The point of making this point is to emphasize that there will never be a universe in which resources have a “unique identifier” in the sense of only having one.”

Through rich linkages with complementary data from trusted sources, libraries can increase the value of their own data beyond the sum of their sources taken individually.

By using globally unique identifiers to designate works, places, people, events, subjects, and other objects or concepts of interest, libraries allow resources to be cited across a broad range of data sources and thus make their metadata descriptions more richly accessible. The Internet's Domain Name System assures stability and trust by putting these identifiers into a regulated and well-understood ownership and maintenance context. This notion is fully compatible with the long-term mandate of libraries. Libraries, and memory institutions generally, are in a unique position to provide trusted metadata for resources of long-term cultural importance as data on the Web.

Another powerful outcome of the reuse of these unique identifiers is that it allows data providers to contribute portions of their data as statements. In our current document-based ecosystem, data is exchanged always in the form of entire records, each of which is presumed to be a complete description. Conversely, in a graph-based ecosystem an organization can supply individual statements about a resource, and all statements provided about a particular uniquely identified resource can be aggregated into a global graph. For example, one library could contribute their country's national bibliography number for a resource, while another might supply a translated title. Library services could accept these statements from outside sources much as they do today when ingesting images of book covers. In a Linked Data ecosystem, there is literally no contribution too small -- an attribute that makes it possible for important connections to come from previously unknown sources.

Library authority data for names and subjects will help reduce redundancy of bibliographic descriptions on the Web by clearly identifying key entities that are shared across Linked Data. This will also aid in the reduction of redundancy of metadata representing library holdings.

Benefits to researchers, students, and patrons

It may not be obvious to users of library and cultural institution services when Linked Data is being employed because the changes will lie "under the hood." As the underlying structured data becomes more richly linked, however, the user may notice improved capabilities for discovering and using data. Navigation across library and non-library information resources will become more sophisticated. Federated searches will improve through the use of links to expand indexes, and users will have a richer set of pathways for browsing.

Linked Data builds on the defining feature of the Web: browsable links (URIs) spanning a seamless information space. Just as the totality of Web pages and websites is available as a whole to users and applications, the totality of datasets using RDF and URIs presents itself as a global information graph that users and applications can seamlessly browse by resolving trails of URI links ("following one's nose"). The value of Linked Data for library users derives from these basic navigation principles. Links between libraries and non-library services such as Wikipedia, Geonames, musicbrainz, the BBC, and The New York Times will connect local collections into the larger universe of information on the Web.

Linked Data is not about creating a different Web, but rather about enhancing the Web through the addition of structured data. This structured data, expressed using technologies such as RDF in Attributes (RDFa) and microdata, plays a role in the crawling and relevancy algorithms of search engines and social networks, and will provide a way for libraries to enhance their visibility through search engine optimization (SEO). Structured data embedded in HTML pages will also facilitate the re-use of library data in services to information seekers: citation management can be made as simple as cutting and pasting URIs. Automating the retrieval of citations from Linked Data or creating links from Web resources to library resources will mean that library data is fully integrated into research documents and bibliographies. Linked Data will favor interdisciplinary research by enriching knowledge through linking among multiple domain-specific knowledge bases.

Migrating existing library data to Linked Data is only a first step; the datasets used for experiments reported in a paper and the model used by the authors to process that data can also be published as Linked Data. Representing a paper, dataset, and model using appropriate vocabularies and formalisms makes it easier for other researchers to replicate an experiment or to reuse its dataset with different models and purposes. If adopted, this practice could improve the rigor of research and make the overall assessment of research reports outlined in research papers more transparent for easier validation by peers. (See for instance the Enhanced Publications use case.)

Benefits to organizations

By promoting a bottom-up approach to publishing data, Linked Data creates an opportunity for libraries to improve the value proposition of describing their assets. The traditionally top-down approach of library data -- i.e., producing MARC records as stand-alone descriptions for library material -- has survived in part due to funding considerations and by the lack of an obvious alternative to metadata record-centric

systems for business transactions, inventory control, discovery, navigation, and preservation: libraries have not had the resources needed to produce information at a higher level of granularity, but have quite successfully focused in collaborative ways on one aspect of the information topography, collections of owned information objects (physical books and other information objects with physical formats and lately their digital avatars), and depending upon other actors, mainly secondary publishers, to provide access to other genres and formats. With Linked Data, different kinds of data about the same asset can be produced in a decentralized way by different actors, then aggregated into a single graph.

Collective Linked Data approaches that make more efficient and effective the experiences of end users, scholars and students, among them, in discovering relevant information objects of many, perhaps any, genre or format might generate support for moving to Linked Data methods that additionally account for or help manage business transactions, inventory control, and preservation, also partly accomplished in collaborative ways. Linked Data technology can help organizations improve their internal data curation processes and maintain better links between, for instance, digitized objects and their descriptions. It can improve data publishing processes within organizations even where data is not entirely open. Whereas today's library technology is specific to library data formats and provided by an Integrated Library System industry specific to libraries, libraries *and other cultural institutions along with the industries serving them* will be able to use mainstream solutions for managing Linked Data. Adoption of mainstream Linked Data technology will give libraries a wider choice in vendors, and the use of standard Linked Data formats will allow libraries to recruit from, interact with, and exploit a larger pool of developers.

Linked Data may be a first step toward a "cloud-based" approach to managing cultural information -- one that could be more cost-effective than stand-alone systems in institutions. This approach could make it possible for small institutions or individual projects to make themselves more visible and connected while reducing infrastructure costs.

With Linked Open Data, libraries can increase their presence on the Web, where most information seekers may be found. The focus on identifiers allows descriptions to be tailored to specific communities such as museums, archives, galleries, and audiovisual archives. The openness of data is more an opportunity than a threat. Clarification of the licensing conditions of descriptive metadata facilitates its reuse and improves institutional visibility. Data thus exposed will be put to unexpected uses, as in the adage: “The best thing to do to your data will be thought of by somebody else.”

Benefits to librarians, archivists, and curators

The benefits to patrons and organizations will also have a direct impact on library professionals. By using Linked Open Data, libraries will create an open, global pool of shared data that can be used and re-used to describe resources, with a limited amount of redundant effort compared with current cataloging processes.

The use of the Web and Web-based identifiers will make up-to-date resource descriptions directly citable by catalogers. The use of shared identifiers will allow them to pull together descriptions for resources outside their domain environment, across all cultural heritage datasets, and even from the Web at large. Catalogers will be able to concentrate their effort on their domain of local expertise, rather than having to re-create existing descriptions that have been already elaborated by others.

History shows that all technologies are transitory, and the history of information technology suggests that specific data formats are especially short-lived. Linked Data describes the meaning of data ("semantics") separately from specific data structures ("syntax" or "formats"), with the result that Linked Data retains its meaning across changes of format. In this sense, Linked Data is more durable and robust than metadata formats that depend on a particular data structure.

Benefits to developers and vendors

Library developers and vendors will directly benefit from not being tied to library-specific data formats. Linked Data methods support the retrieval and re-mixing of data in a way that is consistent across all metadata providers. Instead of requiring data to be accessed using library-centric protocols (e.g., Z39.50), Linked Data uses well-known standard Web protocols such as the Hypertext Transport Protocol (HTTP) and widely used publishing mechanisms and protocols, possibly opening .

Developers will also no longer have to work with library-specific data formats, such as MARC, which require custom software tools and applications. Linked Data methods involve pushing data onto the Web in a form that is generically understandable. Library vendors that support Linked Data will be able to market their products outside of the library world, while vendors *presently* outside the library world may be able to adapt their more generic products to the specific requirements of libraries. By leveraging RDF and HTTP, library and other developers are freed from the need to use domain-specific software, opening a growing range of generic tools, many of which are open-source. They will find it easier to build new services on top of their data. This also opens up a much larger developer community to provide support to information technology

professionals in libraries. *In a sea of RDF triples, no developer is an island.* Correspondingly, in an environment with more offerings and more suppliers, one would expect downward pressure on costs to libraries.

The Current Situation

Issues with traditional library data

Library data is not integrated with Web resources

Library data today resides in databases, which, while they may have Web-facing search interfaces, are not more deeply integrated with other data sources on the Web. There is a considerable amount of bibliographic data and other kinds of resources on the Web that share data points such as dates, geographic information, persons, and organizations. In a future Linked Data environment, all these dots could be connected.

Library standards are designed only for the library community

Many library standards, such as the Machine-Readable Cataloging format (MARC) or the information retrieval protocol Z39.50, have been (or continue to be) developed in a library-specific context. Standardization in the library world is often undertaken by bodies focused exclusively on the library domain, such as the International Federation of Library Associations and Institutions (IFLA) or the Joint Steering Committee for Development of RDA (JSC). By broadening their scope or liaising with Linked Data standardization initiatives, such bodies can expand the relevance and applicability of their standards to data created and used by other communities.

Library data is expressed primarily in natural-language text

Most information in library data is encoded as display-oriented, natural-language text. Some of the fields in MARC records use coded values, such as fixed-length strings representing languages, but there is no clear incentive to include these in all records, since most coded data fields are not used in library system functions. Some of the identifiers carried in MARC records, such as ISBNs for books, could in principle be used for linking, but only after being extracted from the text fields in which they are embedded (i.e., "normalized").

Some data fields, such as authority-controlled names and subjects, have associated records in separate files, and these records have identifiers that could be used to represent those entities in library metadata. However, the data formats in current use do not always support inclusion of these identifiers in records, so many of today's

library systems do not properly support their use. These identifiers also tend to be managed locally rather than globally, and hence are not expressed as URIs which would enable linking to them on the Web. The absence or insufficient support of links by library systems raises important issues. Changes to authority displays require that all related records be retrieved in order to change their text strings -- a disruptive and expensive process that often prevents libraries from implementing changes in a timely manner.

The library community and Semantic Web community have different terminology for similar metadata concepts

Work on library Linked Data can be hampered by the disparity in concepts and terminology between libraries and the Semantic Web community. Few librarians speak of metadata "statements," while the Semantic Web community lacks notions clearly equivalent to "headings" or "authority control." Each community has its own vocabulary, and these reflect differences in their points of view. Mutual understanding must be fostered, as both groups bring important expertise to the construction of a web of data.

Library technology changes depend on vendor systems development

Much of the technical expertise in the library community is concentrated in the small number of vendors who provide the systems and software that run library management functions as well as the user discovery service -- systems which integrate bibliographic data with library management functions such as acquisitions, user data, and circulation. Thus libraries rely on these vendors and their technology development plans, rather than on their own initiative, when they want to adopt Linked Data at a production scale.

Library Linked Data available today

The success of library Linked Data will rely on the ability of practitioners to identify, re-use, or link to other available sources of Linked Data. However, it has hitherto been difficult to get an overview of libraries datasets and vocabularies available as Linked Data. The Incubator Group undertook an inventory of available sources of library-related Linked Data (see Appendix A @@@CITE@@@), leading to the following observations.

Fewer bibliographic datasets have been published as Linked Data than value vocabularies and element sets

Many metadata element sets and value vocabularies have been published as Linked Data over the past few years, including flagship vocabularies such as the Library of Congress Subject Headings and Dewey Decimal Classification. Key element sets, such as Dublin Core, and reference frameworks such as Functional Requirements for Bibliographic Records (FRBR) have been published as Linked Data or in a Linked Data-compatible form.

Relatively fewer bibliographic datasets have been made available as Linked Data, and relatively less metadata for journal articles, citations, or circulation data -- information which could be put to effective use in environments where data is integrated seamlessly across contexts. Pioneering initiatives such as the release of the British National Bibliography reveal the effort required to address challenges such as licensing, data modeling, the handling of legacy data, and collaboration with multiple user communities. However, they also demonstrate the considerable benefits of releasing bibliographic databases as Linked Data. As the community's experience increases, the number of datasets released as Linked Data is growing rapidly.

The quality of and support for available data varies greatly

The level of maturity or stability of available resources varies greatly. Many existing resources are the result of ongoing project work or the result of individual initiatives, and describe themselves as prototypes rather than mature offerings. Indeed, the abundance of such efforts is a sign of activity around and interest in library Linked Data, exemplifying the processes of rapid prototyping and "agile" development that Linked Data supports. At the same time, the need for such creative, dynamically evolving efforts is counterbalanced by a need for library Linked Data resources that are stable and available for the long term.

It is encouraging that established institutions are increasingly committing resources to Linked Data projects, from the national libraries of Sweden, Hungary, Germany, France, the Library of Congress, and the British Library, to the Food and Agriculture Organization of the United Nations and OCLC Online Computer Library Center, Inc. Such institutions provide a stable foundation on which library Linked Data can grow over time.

Linking across datasets has begun but requires further effort and coordination

Establishing connections across datasets realizes a major advantage of Linked Data technology and will be key to its success. Our inventory of available data (see Appendix A) shows that many semantic links have been created between published value

vocabularies -- a great achievement for the nascent library Linked Data community as a whole. More can -- and should -- be done to resolve the issue of redundancy among the various authority resources maintained by libraries. More links are also needed among datasets and among the metadata element sets used to structure Linked Data descriptions. Key bottlenecks are the comparatively low level of long-term support for vocabularies, the limited communication among vocabulary developers, and the lack of mature tools to lower the cost for data providers to produce the large amount of semantic links required. Efforts have begun to facilitate knowledge sharing among participants in this area as well as the production and sharing of relevant links (see the section on linking in Appendix B).

Rights issues

Rights ownership is complex

Some library data has restricted usage based on local policies, contracts, and conditions. Data can therefore have unclear and untested rights issues that hinder their release as Open Data. Rights issues vary significantly from country to country, making it difficult to collaborate on Open Data publishing.

Ownership of legacy catalog records has been complicated by data sharing among libraries over the past fifty years. Records are frequently copied and the copies are modified or enhanced for use by local catalogers. These records may be subsequently re-aggregated into the catalogs of regional, national, and international consortia. Assigning legally sound intellectual property rights between relevant agents and agencies is difficult, and the lack of certainty hinders data sharing in a community which is necessarily extremely cautious on legal matters such as censorship and data privacy and protection.

Data rights may be considered business assets

Where library data has never been shared with another party, rights may be exclusively held by agencies who put a value on their past, present, and future investment in creating, maintaining, and collecting metadata. Larger agencies are likely to treat records as assets in their business plans and may be reluctant to publish them as Linked Open Data, or may be willing to release them only in a stripped- or dumbed-down form with loss of semantic detail, as when "preferred" or "parallel" titles are exposed as a generic title, losing the detail required for use in a formal citation.

[This is the end of the quote from WC3.]

COMPARING CLASSIC MARC DATA RECORD TO LINKED DATA APPROACH

Tim Hodson, in his July post *British Library Data Model: Overview*, provides one scan of how Linked Data might be modeled in ways that contrast with the objectives and structure of MARC records. His treatment of a real-life linked-data model helps extend the W3C textual definitions of Linked Data in library settings.

[suggestion: bring a PDF view of the BL model up in a separate browser window:

http://consulting.talis.com/wp-content/uploads/2011/07/British-Library-Data-Model1.pdf]

Hodson's post includes these thoughts:

One of the key concepts of Linked Data is to represent data as a set of interlinked things. These things are referred to as objects of interest, they are things about which we can make statements.

MARC records are full of statements about various objects of interest. There are books, serials, authors, publishers, times when events happened (such as the publishing of a book), subjects, and identifiers. These things are all things about which more can be said.

One of the key questions that helped the British Library Metadata Services team think about their data in a new way was:

"What is the cataloguer holding in their hand when they record the BNB cataloguing data in the MARC record?"

The obvious answer is 'a book' or 'a serial'. The next questions follow from that initial one, and build a picture of what the cataloguer is holding.

Who wrote the book?

When was the book published?

Who published the book?

Where was the book published?

What is the book about?

What language is it written in?

…

Data reuse is at the core of what Linked Data – as an approach – aims to achieve, whether that data is for reuse internally or externally is for the organization to decide. Data reuse is made easier through the self- describing nature of Linked Data. This means that each property used to describe the relationships between two things is itself described using the same data format that describes the data. Therefore a developer wanting to work with a new set of Linked Data, can look at what properties and types of things they will find in the data and begin to navigate the data to find the things that interest them.

It will be noticed that the majority of the model reuses existing properties and classes from descriptive schema that describe the data the British Library is interested in. Where there was not an appropriate class or property, this was described in the British Library Terms (BLT) schema (this will be formally published in the next couple of weeks).

http://consulting.talis.com/2011/07/british-library-data-model-overview/

WORKSHOP PRODUCTS

VALUE STATEMENT: WHY LINKED DATA APPROACHES ARE WORTH PROTOTYPING/MODELING:

During a time-boxed exercise of the workshop participants, four work groups were asked to produce brief statements highlighting the value of the Linked Data approaches. These statements were then consolidated in a facilitated discussion among all four workgroups in a thirty minute exercise that filled in the gaps and de-duplicated the areas of overlaps. These seven points emerged as a consensus statement, and pithy expression of the value of leveraging Linked Data in the library ecosystem.

1. Linked open data (LOD) puts information where people are looking for it – on the web
2. LOD can expands discoverability of our content
3. LOD opens opportunities for creative innovation in digital scholarship and participation
4. LOD allows for open continuous improvement of data
5. LOD creates a store of machine-actionable data on which improved services can be built
6. Library linked open data might facilitate the break down the tyranny of domain silos
7. LOD can provide direct access to data in ways that are not currently possible, and provides unanticipated benefits that will emerge later as the stores of LOD expand exponentially.

Two examples of sites utilizing Linked Data for navigation and discovery purposes are

1. LinkSailor, a Talis experiment

http://linksailor.com/nav

Give mark twain a try … LinkSailor picks up 1900+ citations for his writing of which the first 120 are listed:

http://linksailor.com/nav?uri=http%3A//semanticlibrary.org/people/mark-twain

2. Rural West Initiative at the Bill Lane Center

http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers

This project is a visualization plotting the development of 140,000 newspapers published over three centuries in the United States. The data comes from the Library of Congress' "Chronicling America" project, which maintains a regularly updated directory of newspapers.

Go to bottom of "introduction" and click on VIEW MAP and the timeline at the top then activates the plot. The segment addressing the West Coast between 1849 and 1860 is interesting in that the discovery of gold stimulated the establishment of numerous newspapers. Note as well that construction of the transcontinental railroad began in 1863.

For additional commentary on this topic, also see the accompanying survey at:

http://www.clir.org/pubs/archives/linked-data-survey/part03_why.html

MANIFESTO FOR LINKED LIBRARIES (AND MUSEUMS AND ARCHIVES AND…)

Based on the early experiences of the workshop participants in the Linked Data ecosystem, and long histories in the libraries and cultural heritage institutions, the workshop participants recognized several typical stumbling blocks that can threaten to trip up progress in both library and other Linked Data initiatives. The participants recognized that to foment the development of a disruptive paradigm for knowledge representation and discovery on the web, the library community will need to depart from "doing business as usual" and adopt new psychologies and new approaches to both metadata and collaboration. A working session among all the workshop participants produced a "Manifesto for Linked Libraries (*et al.*)", consciously patterned after the Agile Manifesto (http://www.agilemanifesto.org). The early Agile software development movement is in many ways similar to the current linked library movement, as an *avant-garde* of practitioners looks to define a new model of productivity in sharp contrast to a "tried and true", but structurally constrained, approach.

We in the cultural heritage and knowledge management institutions are discovering better ways of publishing, sharing, and using information by linking data and helping others do the same. Through this work, we have come to value and to promote the following practices:

1. Publishing data on the web for discovery and use, rather than preserving it in dark, more or less unreachable archives that are often proprietary and profit driven;
2. Continuously improving data and Linked Data, rather than waiting to publishing "perfect" data;
3. Structuring data semantically, rather than preparing flat, unstructured data;
4. Collaborating, rather than working alone;
5. Adopting Web standards, rather than domain specific ones;
6. Using open, commonly understood licenses, rather than closed and/or local licenses.

While we recognize the need for both approaches in each "couplet", we value the initial ones more.

On point 2, some participants in the Workshop asserted the need for 97% accuracy to instill confidence as opposed to improvement (or not) over time. Other participants asserted that this was an artifact of the current cataloging regime, but not entirely necessary, because:

(1) the constant iteration of improvements in URIs that does and will occur to improve accuracy; and

(2) the relevance of accuracy to any individual user being pointillist, valid for given items/topics of interest, but rarely so over an entire database.

On point 4, there are many cooperative programs out there (PCC, etc.). The whole basis of OCLC is sharing and collaboration. So what are we really saying here? The point here is that the current collaboration is done by a very closely-knit group of cataloging specialists. We hope to expand this collaboration to all the data that members of the academic communities (in our context) and many others (in other contexts) are creating

SEEDING A LINKED DATA ENVIRONMENT FOR LIBRARIES

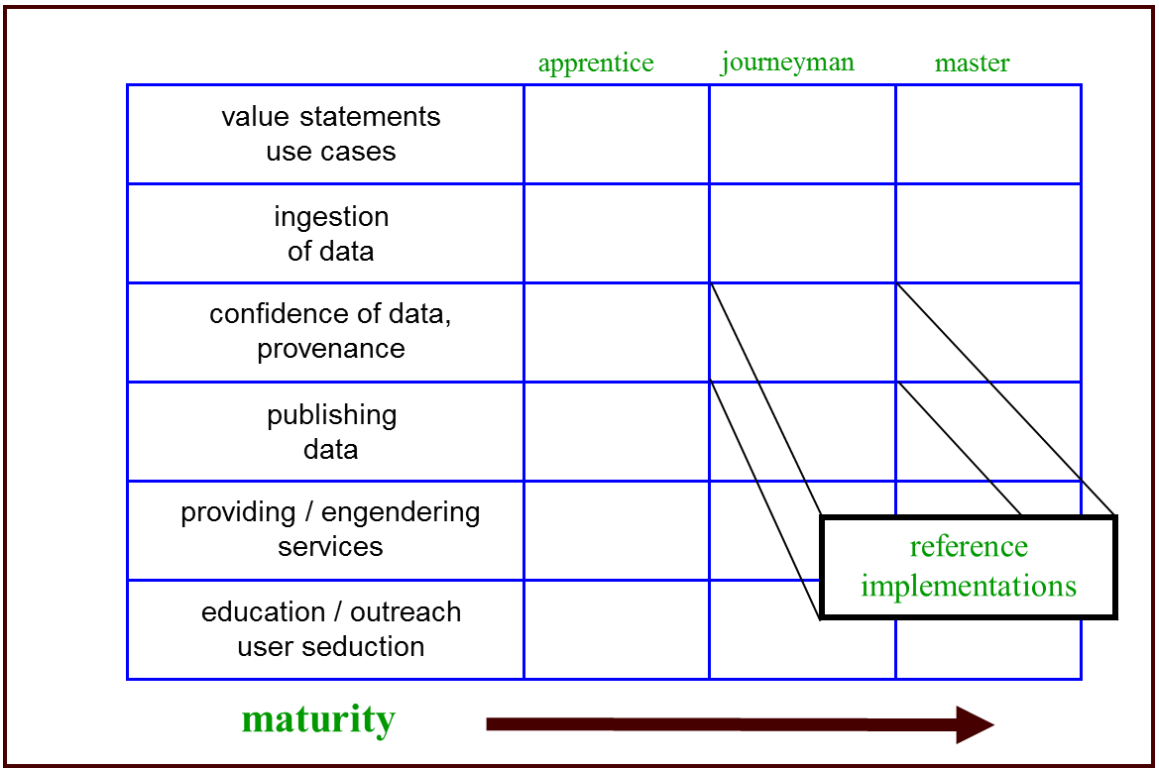
A workflow, principally addressing the transcoding of generic MARC data through an RDF pipeline was identified and presented in diagram form; see Appendix A. Producing a usable and useful Linked Data environment requires generating, using and improving Linked Data stores and services in an iterative approach. These can be described as phases in a life cycle. Those phases proceed from embracing the value proposition of Linked Data approaches and appreciating examples of Linked Data services in operation. The next phases are:

1. constructing use cases;
2. ingesting data (making use of structured data from open stores, constructing or transcoding Linked Data as well as performing quality control);
3. publishing the data, presumably openly so others might use it;
4. providing services based on structured data that is responsive to the use cases;
5. repeating the steps 1. – 5. to add or update use cases, to get new, relevant data, to improve data, and to evolve services;
6. educating producers of metadata (e.g. publishers, librarians, scholarly project leaders) and marketing the resulting services to end users.

A necessary condition is high confidence in the quality of the structured data; some Workshop participants asserted that data not accurate to the level of over 97% produced

user discontent. This assertion needs testing and needs as well to be seen in the context of constant iteration and improvement in URI accuracy as many services contribute and use Linked Data to the wild.

The workflow in Appendix A could provide the basis for the tutorials, etc. referenced above. We started to explore a matrix approach – based on the Linked Data maturity of the institution and the phase of the lifecycle they were trying to achieve. The following matrix could be populated with the specific references that would be relevant to a given institutions needs given their maturity/phase.



Reflecting upon the goals participating institutions might achieve through the use of Linked Data as well as providing business and use cases for Linked Data approaches produced the following examples:

1. Achieving Goals:

    a. My organization's mission includes providing leadership & support to the greater library community. A Linked Data project could further that.

b. My organization's mission is to avail all information, for all people, at all times, to be a hub for information in the geography and culture of the institution's region, and provide this for free. Linked Data has the promise of enabling a richer hub for information distribution. Integration of information for disparate audiences *is* a core mission.

c. My organization's goals are to provide leadership and assistance in information management in a broader organizational context, for schools, departments and research groups over which we have no control. Linked Data may provide a lower-risk, incremental, evolutionary approach to enabling information management.

2. What are the business cases for using Linked Data approaches?

   a. Data integration is easier.

   b. Researcher burden for information sharing, discovery, & reuse is reduced

   c. There is better, faster, cheaper information management.

   d. There is multilingual support.

   e. There is better exposure of institutional resources, thus increasing institutional reputation.

   f. Exposing metadata in Linked Open Data makes more apparent to the public the value of holdings and services of cultural heritage organizations.

3. What are the use cases?

   a. Use Linked Data to streamline authority control:

      eliminate batch processes through obscure logic done externally. Authority control via Linked Data is more immediate, internationalized, more transparent, and more in control. It also enables authority control at point of entry (e.g. by depositor, by producer).

   b. Data globalization:

      Patent research is one example. I'm a researcher and want to see most recent patents happening in photo-voltaics across the world, including in non-English speaking countries. Move to relations via URI's, not labels; URIs provide actionable relationship statements.

c. Connect institutional holdings

Connect holdings as reflected in EADs and similar to related resources from other meta-data created by other institutions; connect local resources to holdings elsewhere, saving time and effort, while exposing more of the documentation, information, knowledge, and/or artifactual record across several or many cultural heritage institutions to scholars and students.

d. Correlate geo-spatial information

Information that could be better transmitted using GIS techniques across the extremely wide variety of textual, image, and quantitative data genres and formats.

PRIORITIZED LIST OF KNOWN ISSUES

Earlier work has shown some specific technical, social and integration challenges in utilizing Linked Data stores and services at scale in the library community. The workshop participants produced a rank-ordered shortlist of specific challenges that the library and cultural heritage community must address for Linked Data to provide a viable solution to the specific needs and challenges of our domain.

Though our list was ranked across categories, many items on the list fall into four major categories:

- Provenance
- Usability
- Preservation
- Standards

The list below shows categories, where applicable, in parenthesis.

1. Cross format referencing, co-referencing, reconciliation (Standards)

   This is an area in which there is much activity, some of it involving efforts to create and promote standardized means of stating relationships between data statements. Much of what is going on tends to be ad hoc experimentation by people working on projects and needing to make statements about connections despite there being many consistent standardized tools and bodies of practice for doing so.

At one level the issues involve two statements of fact and making a decisions about whether the two statements are identical (owl:sameAs in linked-data parlance)[4]. As an example, the question might be whether the URI in one Dublin Core dc:creator statement is equivalent to the URI in another dc:creator statement minted by a different agency (or by a different process within the same agency).

At a second level of complexity, one must deal with range of structured data vocabularies when planning to create, publish, and manage Linked Data. Taking the library convention of a personal author stored in a MARC 100 field as one simple example, there are many ways to represent that name as structured data. With each of those alternatives treating such a name within a different semantic context, owl:sameAs finds itself being used as a bridge between vocabularies when the strict sense of "sameAs" may not apply.

This call for participation in a DCMI-2011 Special Session on Vocabulary management and alignment summarizes some of the issues that are in play:[5]

"Agenda: At DC-2010, Mike Bergman's keynote strongly suggested that DCMI has a potential role in promoting co-operation among vocabulary managers and in providing best practices for vocabulary alignment and interoperability. The inevitable and useful proliferation of vocabularies emerging in the Linked Data space demonstrates a need for increased vocabulary reuse and tools to facilitate this reuse, as well as central reference vocabularies and tools to manage and encourage vocabulary mapping. Recent announcements about search engine support for schema.org and microdata make the need even more prominent.

Toward this end, a first step was recently taken with the announcement of a collaboration effort between DCMI and FOAF. This full day special session will explore the scope and nature of vocabulary management issues, with illustrations from a variety of different domains and communities, and discuss a variety of proposals and ideas for how DCMI, the W3C and other committed organizations might contribute to both infrastructure and best practices for more effective vocabulary management and interoperability. For the purpose of this workshop, "vocabularies" refers to both property / element sets and value / controlled vocabularies."

---

[4] OWL = WC3 Web Ontology Language; see http://www.w3.org/TR/owl-ref/
[5] Dublin Core Metadata Initiative; see http://dublincore.org/

http://dcevents.dublincore.org/index.php/IntConf/index/pages/view/specialSessions-2011The notes associated with GRP#1 workflow are pertinent here.

Sameas.org provides a brief set of citations for this topic

2. Use of library authority files – names, subjects, etc. (Standards and Usability)

   Library metadata is an excellent first source for Linked Data in part because of the authority files that support its controlled terms. The records of authority files can be readily published with stable, persistent URIs and the data within those records – variant terms and relationships to other terms – are valuable for broader matching. However, while the relationships between one authority and others may be expressed in a programmatic way, the related terms are entered into present-day records only as lexical strings. Although many authority files have been very successfully published as Linked Open Data, it has taken considerable programming to disambiguate and match those lexical strings with their unambiguous identifiers.

   As the cataloging community moves to adopt RDA and embrace Linked Data, the evolution of authority files must keep pace. With the ability to control headings by a direct link to the heading's URI, unique text strings for each separate heading in the authority file are no longer required. Moreover, by using HTTP URIs, additional information about a concept or name is readily accessible.

   Numerous authority files, standards, and registries exist to support particular functions in regard to the identification and control of names, subject headings, and other value vocabularies. ORCID and MIMAS are developing author registries to control attribution in journal literature; the International Standard Name Identifier (presently a draft ISO standard for the identification of public identities of parties) provides a means to generate a unique identifier for someone or something with a public identify; and traditional national authority files (such as the Deutche Nationalbibliothek's name authority file) contain millions of carefully curated entries of personal and corporate names. A particular heading might appear in any or all of these files in a variety of forms. Continuing work in this arena is needed, including the publication of more open authority data. Ultimately, by linking parallel URIs in all of these sources, a powerful web of associations can be created that will dramatically benefit the accuracy of machine-generated links.

   The open licensing of authority data – preferably either by pushing the data into the public domain or publishing data with a Creative Commons CC0 license – is vital

since it empowers data consumers with the freedom to use, re-use, link to, and otherwise re-purpose the data to best fit their particular needs.

Although not strictly a Linked Data issue, the rules for creating the unique strings used as subject headings and names are quite complex and severely limit the number of people qualified to create them. Because only a relative few individuals are able to create authorized headings and because URIs are the preferred method of "authority control" in the LD realm, this has impeded the process of creating URIs for a good many subject heading terms and bibliographic identities. It is time to (re-)evaluate the value of these precise strings to the overall description of library resources and to how users search for and utilize this information. Treating authority information as "data," versus a controlled string, can lead to refined faceting of the information and improved display and discovery. A larger discussion must take place between Linked Data practitioners and those who create authority records about how present data formats and technologies can enhance the search and discovery experience, but which may be impeded by current cataloging rules and best practices.

In late September 2011, the Conference of European National Librarians (CENL) made a bold statement endorsing the open licensing of their bibliographic data. Of equal or greater importance will be the open licensing of their authority data. The authority files support the controlled headings in the bibliographic data they have made available. Without them, the interlinking of this data will be severely hampered.

3. Killer app(s) (Usability)

In retrospect, we should have better defined this category. It was a source of significant discussion, but has different meanings for different individuals. One concept suggested was a multi-institution map project. Another suggestion from was the Civil War 150 website. Imagine being able to (automatically) populate a website that could allow users to navigate through Civil War history from different perspectives – all from Linked Data. One could explore events based on time, place, person, etc. Not just faceted browsing, but an interactive experience.

There have been glimmers of development in this area, but nothing that steps out at a clearly new level of search or navigational capabilities. Here are some tantalizing examples in miniscule of the possibilities, ones that may lead to the development of more comprehensive environments for discovery and navigation based on new user interfaces working on large stores of Linked Data records.

- David Huynh (MIT) has produced a video overview of his prototype parallax in 2008. http://vimeo.com/1513562
- The BBC's wildlife sub-site is all driven by linked data under the hood. Richard Wallace summarizes the site's features in his presentation at the British Library in July:
    - slides 63-75 http://www.slideshare.net/rjw/linked-data-applicable-for-libraries
    - minutes 51:45 -- 55:45 in video http://www.ustream.tv/recorded/15986081
    - NOTE: related to the last 2 slides, note that the BBC wanted to add dinosaurs to the wildlife site, a significant task in most database environments. The effort was completed in a couple of days by extending the ontologies behind the linked data in the BBC site.
- LinkSailor, a Talis experiment (http://linksailor.com/nav)
    - Searching Mark Twain picks up 1900+ citations for his writing of which the first 120 are listed:
    - http://linksailor.com/nav?uri=http%3A//semanticlibrary.org/people/mark-twain
- Maybe the most comprehensive "showy/eye-catching" example of what could be done with linked data is the Civil War 150 site (http://www.history.com/interactives/civil-war-150#/home). The site's access that cuts across all manner of resources (library, archive, museum, visual, textual, graphic, maps, etc.). It provides 25 varied facets for access to the details under headings for
        - Technology
        - Union
        - Confederate
        - Battles
        - Places/Events
        - Culture
    - See technical commentary at http://radar.oreilly.com/2011/04/linked-data-civil-war.html and http://www.civilwardata150.net/news/

4. Attribution, origin, & authority (Provenance)

   This was a recurring theme. The provenance of data seems to be one of the biggest challenges we face in the Linked Data world. It underscores the balance we have to achieve between openness (not waiting for something to be perfect) and accuracy. Training in the creation, derivation, and publication of URIs, as well as making links, and using links in discovery environments (Usability)

   It is also partially about people understanding that the URI has the attributes implied by the publisher (which is always the domain owner). This is a very strong fact, contrasting with a common perception of the web, which is the opposite, in which there is no strong ownership of a URL by the originator. There is also a technical issue of how to represent these attributes, currently a topic of active research.

5. Training in the creation, derivation, and publication of URIs, as well as making links, and using links in discovery environments (Usability)

   Creating and publishing URIs is not a difficult technical problem set. The hard technical questions are those around "reification" and the expression of metadata.

6. Usability of data (Usability)

   Data must be "reificate-able". The ability to specify properties such as trust and provenance of RDF data requires the system to be able to make metadata statements (the trust statement) about the metadata (the RDF, such as a catalogue record). This ability to consider the data itself as a Thing to be referred to is known as "reification".

   The general issue of reflecting reification from the logic underlying RDF to implementations is still a topic of active research.

   However, in practice almost all RDF systems provide sufficient technology to enable properties of trust etc. to be asserted and accessed, usually by the use of an extension to the RDF store and associated SPARQL known as Named Graphs.

   See also the 'Statement reification and context in the Wikipedia article on Resource Description Frameworkhttp://en.wikipedia.org/wiki/Resource_Description_Framework)

7. Quality Control (Usability)

QC must be both accomplished as URIs are created and performed iteratively over time. QC of unfamiliar languages, either for metadata or information, is a special challenge.

8. Standards for URIs (Standards)

Together, Kyle Neath and Jeni Tennison provide a thorough survey of URL design.

For our purposes here, we can safely equate URLs and URIs … one of the chief tenets of Linked Data (per Tim Berners Lee) is that URIs must be resolvable, and more importantly, when they resolve they should point to useful information.

You should take time to design your URL structure. If there's one thing I hope you remember after reading this article it's to take time to design your URL structure. Don't leave it up to your framework. Don't leave it up to chance. Think about it and craft an experience.

URL Design is a complex subject. I can't say there are any "right" solutions — it's much like the rest of design. There's good URL design, there's bad URL design, and there's everything in between — it's subjective.

But that doesn't mean there aren't best practices for creating great URLs. I hope to impress upon you some best practices in URL design I've learned over the years …

Why you need to be designing your URLs

Top level sections are gold

Name spacing is a great tool to expand URLs

Query strings are great for filters and sorts

Non-ASCII URLs are terrible for English sites

URLS are for humans – not for search engines

A URL is an agreement

Everything should have a URL

A link should behave like a link

Post-specific links need to die

Kyle Neath

9. Data Curation (Preservation)

Linked Data uses URIs. Linked Data can thus be collected for preservation by archives other than the original publisher using existing web crawling techniques

such as the Internet Archive's *Heritrix*. Enabling multiple archives to collect and preserve Linked Data will be essential; some of the publishers will inevitably fail for a variety of reasons. Motivating web archives to do this will be important, as will tools to measure the extent to which they succeed. The various archives preserving Linked Data items can republish them, but only at URIs different from the original one, since they do not control the original publisher's DNS entry. Links to the original will not resolve to the archive copies, removing them from the world of Linked Data. This problem is generic to web archiving. Solving it is enabled by the *Memento* technology, which is on track to become an IETF/W3C standard. It will be essential that both archives preserving, and tools accessing Linked Data implement Memento. There are some higher level issues in the use of Memento, but as it gets wider use they are likely to be resolved before they become critical for Linked Data. Collection using web crawlers and re-publishing using Memento provide archives with a technical basis for linked open data preservation, but they also need a legal basis. Over 80% of current data sources do not provide any license information; these sources will be problematic to archive. Even those data sources that do provide license information may be problematic, their license may not allow the operations required for preservation. Open data licenses do not merely permit and encourage re-use of data, they permit and encourage its preservation.

10. Distribution of responsibility (Usability)

This heading came to cover a varied collection of topics as the workshop carried forward. Included were:

a. Preservation of data ... this is addressed under 9. Data Curation, above.

b. Feedback, reporting, reward systems, metrics, motivation for contributing Linked Data and or/URIs

c. Gaming and competition ... this is addressed under 11. Marketing/Outreach, below.

With respect to item *b. Feedback ... ,* the very nature of Linked Data lends little to the pursuit of measuring benefits through statistics and other types of objective metrics. Having created a pool of Linked Data and made it openly available for use on the web, there are few tools that can see how and by whom that newly released data is

being used. Uniquely formed URIs might be traceable in some manner, and restrictions like CC-BY might generate some feedback events. Too, those URIs for which an organization is the only (or the primary) resolving agency do have a means to measure that resolution traffic.

In general, however, it may well be that in order to take advantage of emergent semantic-web capabilities in/on the web, organizations will need to take a strategic decision that they should (must?) contribute to the scope/density of emerging linked-data environs. This, because moving up the learning curve for creating and sharing Linked Data, may in fact be the most effective way to acquire the knowledge and experience that allows an organization to effectively exploit emerging forms of structured, web-wide data as the evolution of structured data toward Linked Data and beyond that toward future forms of semantic data continues. Here the investment is contributing to the scope and density of links and the ROI is capabilities that allow an organization to exploit that portion of the linked-data-driven web that their efforts have helped to expand and enrich.

11. Marketing/Outreach (Usability)

User seduction & training of staff as well as users are key here.

Also, many types of programs and activities show evidence of being productive in helping advance the uptake of various types of new technologies. For example gaming and competitions have taken various forms.

One example is *Games for Change* (http://en.wikipedia.org/wiki/Games_for_Change):

a global advocate for supporting and making games for social impact. It brings together organizations and individuals from the social impact sector, government, media, academia, the gaming industry, and the arts to grow the field. incubate new projects, and provide an open platform for the exchange of ideas and resources.

Crowdsourcing is another facet of social interaction over the web ... the ubiquitous example being Wikipedia. See also the accompanying survey for some additional sources of information at: http://www.clir.org/pubs/archives/linked-data-survey/part11_c_tools.html .

In terms of rewards, here is an example of a very direct approach:

NYC BigApps 3.0 offers $50,000 in cash and other prizes to software developers for the best new apps that utilize NYC Open Data to help NYC residents, visitors, and businesses. BigApps 3.0 continues New York City's ongoing engagement with the software developer community to improve the City, building on the first two annual BigApps competitions through new data, prizes, and resources. http://2011.nycbigapps.com/

An example of another type of marketing and outreach is a growing grass-roots effort that is currently underway in the linked-open-data--Library/Archive/Museum arena known by its acronym as LOD-LAM. Launched in June at an "un-conference" in San Francisco, it has generated an increasing amount of activity with events and online conversations spread around the US and overseas. One can review the launch and the ongoing project via

- the home web site http://lod-lam.net/summit/ ;
- an introductory video http://lod-lam.net/summit/2011/09/15/intro-to-lodlam-talk-live-from-the-smithsonian/ ;
- its Google Group http://groups.google.com/group/lod-lam ;
- various reading lists http://lod-lam.net/summit/2011/04/25/lodlam-reading-lists/ ; and
- its *About* page http://lod-lam.net/summit/about/ .

12. Workflow (Usability)

The accompanying Literature Survey includes a brief section on Workflows. It includes summaries of a pair of posts by Mike Bergman in which he addresses the need for structure in the face of sometimes overwhelming pressures for simplicity. He refers to a "semantic sweet spot" as his target for an appropriate balance between fully marked-up content and quick-pass solutions. Other viewpoints are included. The overall emphasis in this section of the survey is on "what" might need consideration in relation to planning workflows, rather than the nuts-and-bolts for sequencing of appropriate processes and data flows.

The group recognized that identity management is a crucial part of such a workflow. The group was informed by the presence of Hugh Glaser of Seme4 Limited, who is the creator of http://sameas.org/ and similar services. The generic sameAs service already offers facilities for canonization, deprecation and partition, and it was recognized that these were exactly the sort of facilities that a workflow such as this requires.

A sample workflow for minting URIs, then iteratively reconciling them is shown in Appendix A, below; this is a work product of Working Group 1 in the Workshop.

13. Scalability

    Some participants in the LDW asserted that web scaling is already accomplished and that a natural process of exploitation of URIs will occur spontaneously. Therefore, the challenge is to convert and manufacture URIs, then place them in open stores, and then let whatever interfaces or killer apps there are or will be make use of the open stores of URIs. For a fuller explanation of web scale see: http://community.oclc.org/engineering/2009/05/what-is-web-scale.html .

14. Indexing

    Like Killer Apps, indexing on the basis of URIs and indexing URIs do not yet demonstrate the precision and reliability of results that we now get from word indexing in closely managed pools of metadata. Indexing URIs, as demonstrated by Sindice (http://sindice.com/search) produces large results that presently cannot be refined easily or, as in the case of Freebase (http://www.freebase.com/), produce results that are obviously fragmentary in most categories. However, each of these examples demonstrate the principle of Linked Data approaches that ignore format and genre boundaries and thus show the range of possibilities for improved discovery and navigation. A differentiator of searches based in a Linked Data environment, so far, is the relevance of results on the one hand and the formatting of results in some other cases, e.g. Freebase, based on a chosen schema that displays results for many kinds of information objects in immediately useful ways, i.e. in categories of information, not merely lists of web sites of potential interest. The Web indexing services display information from web sites based on some, usually only partially understood, filters, but without understanding or allowing refinement or presentation by nature of the underlying information object.

15. Use of ontologies (Standards)

    Ontologies, formal representations of concepts within a domain and their relationships to each other, have long been used to organize topics contained in information resources. Full text searching is often inadequate as concepts can be expressed in many semantic variations and in many different languages. By making these ontologies available as linked-data, the concepts within them can be applied consistently and freely across temporal and physical borders. By linking concepts across ontologies for different domains, extremely powerful, automated subject matching is created and a wealth of data retrieved from outside a patron's primary field of research.

The term "ontology" is often used ambiguously to refer to:

1) advanced metadata models, such as CIDOC CRM;

2) domain specific thesauri-like vocabularies listing typically general concepts or classes (thesauri, classifications, subject headings, etc.); and

3) registries of individuals (authority files, geographical gazetteers, event repositories, etc.).

In 1) a major challenge posed to the library Linked Data community is how to align different metadata models used for different kind of library and cultural heritage objects and descriptions intangible phenomena, such as events, into an interoperable collection of Linked Data. A major problem in 2) is how map the different vocabularies used in different domains, disciplines, and cultures with each other to facilitate e.g. query expansion across vocabulary boundaries. Registries 3) pose the library Linked Data community still another set of challenges. The problems of dealing with authority files, e.g. disambiguating between persons with similar names and dealing with the multitude of names and their transliterations in different language, are already well-appreciated in libraries. Similar problems are encountered e.g. when dealing with places, and especially when taking into account historical places that have changed over time. All these issues have to be dealt with on an expanding international level, involving Linked Data coming from different countries, practices, cultures, and in different languages.

See: Eero Hyvönen: **Semantic Portals for Cultural Heritage**. *Handbook on Ontologies (2nd Edition)* (Steffen Staab and Rudi Studer (eds.)), Springer-Verlag, 2009. http://www.seco.tkk.fi/publications/2009/hyvonen-portals-2009.pdf

16. Licensing (Standards)

Questions about licensing metadata are myriad and complex, when such licenses exist and are documented, referenced, or even implied. Organizations like the Open Knowledge Foundation and related efforts/groups are waging what appears to be an increasingly successful campaign to open up metadata under what are dubbed "Creative Commons 0" licenses—any type of [re-]use for any purpose, regardless of commercial or other intent. Witness the recent vote by European National Libraries to open up their metadata.

Meeting at the Royal Library of Denmark, the Conference of European National Librarians (CENL), has voted overwhelmingly to support the open licensing of their data. CENL represents Europe's 46 national libraries, and are responsible for the massive collection of publications that represent the accumulated knowledge of Europe.

https://app.e2ma.net/app/view:CampaignPublic/id:1403149.7214447972/rid:48e64615892ac6adde9a4066e88c736c . This was reported 28 September 2011.

The accompanying Literature Survey includes a scan of the intellectual property landscape in this venue.

17 Annotation (Provenance)

Taken in one way, annotation can be taken as the process of adding commentary to extant content. Such additions might range from a simple personal note to full-scale critical commentary on a complex set of issues and resources. The accompanying Literature Survey provides a two-part introduction to this topic, a look at a project for the academic community, and commentary by a long-time web development pundit.

Taken another way, annotation can be taken as the process of extending and refining (and even debating vagaries of) metadata and other navigational aides to discovering and exploring cultural heritage resources. The Survey provides an extended introduction to this type of activity under the general rubric of crowdsourcing. Included are an ACM Communications' study of the topic, Mark Ockerbloom's summary from a library perspective, and dozen examples from various environments.

18. Identity Management

The workflow presented in Appendix A makes extensive use of an identity management subsystem, of the sort provided at sameAs.org by Hugh Glaser (see Note 6), which is in fact sometimes used by FreeBase. In addition, gaining value from multiple organizations publishing as Linked Data requires identity management that crosses institutional boundaries. During the workshop Hugh brought up a proof of concept site (http://sameas.org/store/kelle/) to show a little of what can be done, solely for subject headings. This has been continued, and since enhanced with other data.

Hugh, on behalf of Seme4, offered to support the LDW activities as best he can, by providing identity management systems for institutions, and for cross-institution activity and projects.

19. Relationship to e-scholarship (esp. e-science) & e-learning

    The proliferation of separable elements (e.g. graphics including photographic images and supplemental data including videos and spreadsheets) attached to or embedded in scholarly communications, particularly articles, since the advent of web publishing in the mid-1990s, as well as analogous elements of courses supported by web-based course management systems suggests a need for much more metadata generation and indexing than previously imagined. That some Internet publishing services, such as HighWire Press, have made easy the downloading to presentation slide sets for papers and class lectures of graphics proves the point. And yet, because of the inherent investment of labor necessary to create metadata compatible with the various indexing, discovery, and navigation systems or schemes operating today, these elements must be discovered through indirect means. Linked Data approaches, optimally generated algorithmically as articles, are processed by publishers and/or their Internet service providers could make separable elements discoverable and ideally save researchers and instructors time and effort. Combining Library Linked Data with Publisher Linked Data and Linked Data from a variety of other sources, including scholarly projects, could lead to dramatically improved discovery and navigation in speed, relevance, and the means for refinement of searches. In addition, that same metadata expressed in Linked Data format could become the underpinning for systems supporting the business operations of libraries, museums, publishers, scholarly institutions and societies, among others.

20. Cultural diversity (Usability)

    One major promise of Linked Data is its inherent compatibility with multilingualism. By representing entities and concepts through URI's rather than text strings, the research and cultural heritage community may be able overcome the stumbling blocks that have tripped up libraries (and others) in searching for relevant information across text bases spanning different languages and character sets. By labeling the same entity with different text stings, linked-data-powered systems can simultaneously support cross-language queries, computation and results retrieval, while presenting results in a user interface that invokes the correct set character strings / translation labels that are appropriate to the user and context. An

international Linked Data environment must, from the outset, factor this internationalization into its design. This includes UI's that can input and output appropriately internationalized strings and displays; it may also include support for schema that can reflect and relate different cultural understandings and contexts for common entities.

21. Search engine optimization (Standards)

The current iteration of structured data, also known as micro-data, aimed at providing better search results (and some would say optimizing the rank of hits on the data offered up by content providers) is best seen at schema.org. Other iterations of related approaches have included Google's "rich snippets" and the linked-data community's offering, RDFs.

From its homepage: **What is schema.org?**

> "This site provides a collection of schemas, i.e., html tags, that webmasters can use to markup their pages in ways recognized by major search providers. Search engines including Bing, Google and Yahoo rely on this markup to improve the display of search results, making it easier for people to find the right web pages.
>
> Many sites are generated from structured data, which is often stored in databases. When this data is formatted into HTML, it becomes very difficult to recover the original structured data. Many applications, especially search engines, can benefit greatly from direct access to this structured data. On-page markup enables search engines to understand the information on web pages and provide richer search results in order to make it easier for users to find relevant information on the web. Markup can also enable new tools and applications that make use of the structure.
>
> A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. So, in the spirit of sitemaps.org, Bing, Google and Yahoo! have come together to provide a shared collection of schemas that webmasters can use."
>
> URL http://schema.org/

In practical terms, one can see micro-data in action in the HighWire Press interface:

HighWire
Stanford University

Info for Authors | Editorial Board | About | Subscribe | Advertise | Contact | Feedback | Site Map

PNAS
Proceedings of the National Academy of Sciences of the United States of America

Heterotrimeric G protein $\beta_1\gamma_2$ subunits change orientation upon complex formation with G protein-coupled receptor kinase 2 (GRK2) on a model membrane

Andrew P. Boughton[a,1], Pei Yang[a,1], Valerie M. Tesmer[b], Bei Ding[a], John J. G. Tesmer[b,1], and Zhan Chen[a,1]

+ Author Affiliations

Edited* by Gabor A. Somorjai, University of California, Berkeley, CA, and approved July 25, 2011 (received for review May 24, 2011)

Abstract

Few experimental techniques can assess the orientation of peripheral membrane proteins in their native environment. Sum Frequency Generation (SFG) vibrational spectroscopy was applied to study the formation of the complex between G protein–coupled receptor (GPCR) kinase 2 (GRK2) and heterotrimeric G protein $\beta_1\gamma_2$ subunits (Gβγ) at a lipid bilayer, without any exogenous labels. The most likely membrane orientation of the GRK2–Gβγ complex differs from that predicted from the known protein crystal structure, and positions the predicted receptor docking site of GRK2 such that it would more optimally interact with GPCRs. Gβγ also appears to change its orientation after binding to GRK2. The developed methodology is widely applicable for the study of other membrane proteins in situ.

Footnotes

↵ [1]A.P.B. and P.Y. contributed equally to this work.

« Previous | Next Article » Table of Contents

This Article

Published online before print August 29, 2011, doi: 10.1073/pnas.1108236108
PNAS September 13, 2011 vol. 108 no. 37 E667-E673

Author Summary
Author Summary (PDF)
Author Summary Figure Only
» Abstract
Figures Only
Full Text
Full Text (PDF)
Full Text + SI (Combined PDF)
Supporting Information

- Classifications
PNAS Plus
Physical Sciences
Chemistry
Biological Sciences
Biophysics and Computational Biology

- Services
Email this article to a colleague
Alert me when this article is cited
Alert me if a correction is posted
Similar articles in this journal
Similar articles in PubMed
Add to My File Cabinet
Download to citation manager
Request copyright permission

+ Citing Articles
+ Google Scholar

Search PNAS GO
advanced search »

This Week's Issue
September 20, 2011, 108 (38)

From the Cover
- Giant gypsum crystals
- Investigating the Black Death
- Game theory and climate policy
- Ancient seed-eating birds
- Gut bacteria affect mouse brains

Alert me to new issues of PNAS

Early Edition
Archives
Online Submission
Feature Articles
PNAS Plus
Commentaries
Letters

14. extracted from tagging within the HTML display page:

itemType: http://schema.org/ScholarlyArticle

15. **articlebody:** Heterotrimeric G protein β1γ2 subunits change orientation …

16. contributor-list

17. contributor-1 author Andrew P. Boughton
18. contributor-2 author Pei Yang
19. …

20. affiliation-list

21. aff-1 Department of Chemistry, University of Michigan
22. aff-2 Life Sciences Institute and the Department of Pharmacology, U of Michigan
23. **abstract-1** Few experimental techniques can assess the orientation of peripheral membrane
24. proteins in their native environment. Sum Frequency Generation (SFG) vibration

25. fn-supplemental material

26. This article contains supporting information online at <a href="/lookup/suppl/doi:10.1073/…

22. Social Media: FaceBook apps and similar

Facebook's Open Graph protocol takes aim at a specific target: providing enough information to represent any web page within the social graph. OGP provides web developers with a framework for adding metadata for four properties to web pages.

This and other "graphed components" of the ever expanding social-media web give linked-data and semantic-web proponents and practitioners considerable pause. Dean Allemang sums up the dichotomy between simple & viral when compared with fully-analyzed & little-used in this summary of his post (Simple, simpler, simplest …?):

From the point of view of metatags, the Open Graph Protocol is really simple; just a handful of required tags with a simplified syntax (simpler even than standard RDFa). Even so, Facebook user studies showed that this was almost too complicated. For some audiences, simple really has to be simple. This is a tough pill for any technologist to swallow; looking at OGP makes it look as if the baby has been thrown out with the bathwater.

But there are now hundreds of millions of new 'like' buttons around the web; simplicity pays off. As another commenter pointed out, regardless of the purity (or lack thereof) of the Facebook approach, OGP has still made the biggest splash in terms of bringing semantic web to the attention of the public at large. So who's the bandwagon, and who's riding?

Conclusion to Issues

For all of these issues, the assignment of assets (staff, outsourcing, money, i.t.) add additional complications.

The most critical realization to come out of the development of this list is the fact that the business case for constructing services in a Linked Data environment must credibly promise improved discovery and navigation for end users.

Those recalling resistance to the retrospective conversion of library card catalogs know how swiftly those with concerns shifted to become avid users of OPACs once recon was considerably along.

DEPLOYING LINKED DATA

Please see the text and diagram in Appendix A.

SEARCHING FOR KILLER APPS

There were numerous invocations of the need for a killer app to demonstrate the validity of the Linked Data approach for discover and navigation across the range of information objects, metadata about them, and even actual objects in cultural memory organizations like libraries, archives, and museums. There have been glimmers, but nothing that steps out at a clearly new level of search or navigational capabilities. Here are some tantalizing examples in miniscule of the possibilities, ones that may lead to the development of more comprehensive environments for discovery and navigation based on new user interfaces working on large stores of Linked Data records.

- David Huynh (MIT) has produced a video overview of his prototype *parallax* in 2008. http://vimeo.com/1513562
- The BBC's wildlife sub-site is all driven by Linked Data under the hood

  The BBC's wildlife sub-site is all driven by Linked Data under the hood. Richard Wallace summarizes the site's features in his presentation at the British Library in July:

- slides 63-75 http://www.slideshare.net/rjw/linked-data-applicable-for-libraries

- minutes 51:45 -- 55:45 in video http://www.ustream.tv/recorded/15986081
    - NOTE: related to the last 2 slides, note that the BBC wanted to add dinosaurs to the wildlife site … a significant task in most database environments – completed in a couple of days by extending the ontologies behind the Linked Data in the BBC site.
- LinkSailor, a Talis experiment

http://linksailor.com/nav

Give mark twain a try … LinkSailor picks up 1900+ citations for his writing of which the first 120 are listed:

http://linksailor.com/nav?uri=http%3A//semanticlibrary.org/people/mark-twain

- Maybe the most comprehensive "showy/eye-catching" example of what could be done with Linked Data is the **Civil Ware 150** site … here's access that cuts across all manner of resources (library, archive, museum, visual, textual, graphic, maps, etc.). It provides 25 varied facets for access to the details under headings for Technology; Union; Confederate; Battles; Places/Events; and Culture

  See: technical commentary at http://radar.oreilly.com/2011/04/linked-data-civil-war.html and http://www.civilwardata150.net/news/

- Metaweb/Freebase, now a Google company, has produced a preliminary model of one way library and web-based Linked Data might be combined to produce a more synoptic view of resources and services for the support of teaching, learning, and research. Go to: http://www.freebase.com ; in the "find topics" box enter any term or name; then click on the numerous results displayed in list form, some with brief descriptive annotations. For more information: http://wiki.freebase.com/wiki/What_is_Freebase%3F .

NEXT STEPS & POTENTIAL PROJECTS

Workshop participants will continue to pursue individual efforts, but expect to collaborate to pursue the goal of advancing Linked Data.

NEXT STEPS

The Stanford team, with the assistance of other participants will generate a model for a multi-national, multi-institutional discovery environment built on Linked Open Data demonstrating to end users, our communities of researchers the value of the Linked Data approach. That model will per force include the basic functions of generating, harvesting, and iteratively reconciling URIs as well as adapting or, if necessary building one or more "killer apps", assembling and/or calling upon tools supporting the necessary steps in the workflow, and then operating the environment for academic information resources. That model will be shared with the participants in this workshop and beyond.

DEFINED PROPOSALS

URI CREATION

Creation of structured data (URIs) from metadata from articles in scholarly journals, a potential joint project between Stanford HighWire Press and the British Library. The target metadata comes from articles running through the HighWire servers (6.7M), metadata from Medline/PubMed (>21M citations), and articles from 20,000 journals for which the British Library has permission to make use of the metadata.

MARC RECORDS

The Stanford team will work with the national libraries represented at the workshop (Library of Congress, British Library, Bibliothèque nationale de France, Deutsche Nationalbibliothek) and others, including research libraries here and abroad. We take heart from the Conference of European National Libraries' (CENL) bold statement in September 2011, voting to support opening up their metadata as linked open data. In that vein, we will follow the lead of the fine work carried out by the British Library, whose staff, in concert with people from Talis: designed a rich, web-savvy data model for library Linked Data; built their Linked Data by extracting appropriate facts from their MARC records; released the data as open data, without constraints on its use.

Our plans also include attention to the various flavors of authority records that underpin today's library metadata (as noted under the VIAF heading that follows).

For an outline of extant meta-data associated with libraries, museums, and archives, also see the accompanying survey at http://www.clir.org/pubs/archives/linked-data-survey/part09_a_metadata.html . Note also that APPENDIX B: provides a scan of sources in table form.

OPEN VIAF

It is hghly desirable to create an "open" VIAF, or requesting OCLC to provide VIAF as an open Linked Data service building on the work of the British Library, the Deutsch Nationalbibliothek, the Royal Library of Denmark, the Library of Congress, the Bibliothèque nationale de France and other institutions' transcoding of name and other authorities to URIs. This "open" VIAF would use Linked Data to streamline authority control, not batch processes through obscure logic done externally. Authority control via Linked Data is more immediate, internationalized, more transparent, and more in control. Also enable authority control at point of entry (by depositor, by producer, e.g.).

MANUSCRIPT INTEROPERABILITY

In the specific domain of digitized ancient, medieval, and early modern manuscripts and in specific support of the work underway to develop the tools and agreements to support interoperability for scholarly functions across silos of digitized manuscripts, Stanford will collect descriptions of manuscripts in URIs. Then, Stanford or another agency will connect individual applications that are showcasing different sets of medieval manuscripts. These projects, the development of interoperability across silos AND the descriptions of manuscripts expressed in URIs, are extensible to many other domains and their digital repositories.

LINKED OPEN DATA TOOL KITS

a. A census of the currently available tools in support of the envisioned prototype and other ones like it is needed.

b. Stores of URIs are readily available, but the tools to generate and use them are hard to find or not existent. There are indicators that this situation may be changing, but there is a problem perceived, because the tools we know about are more generic than what is needed for libraries and other tools are needed for publishers. Rather than

try to build new tools, or use the general ones in the abstract, the provision of configuration tools may be the answer. A question arose: Is it enough to agree to share configurations, and identify it as a future area for engagement?

c. What is the experience of projects making use of the existing tools to create URIs from MARC records?

d. We need tools that are to purpose, not a list of everything that's out there. This is where the idea of a "cookbook" could come in. Tried and tested tools and methodologies that can jumpstart institutions with little or no experience in Linked Data should be provided by a project addressing these needs.

MARC CLEARINGHOUSE

A MARC Clearinghouse (Data Store) should be set up from the URIs derived from the iterative process outlined in Appendix A. The FRBR Group 1 entity relationships between resources should be included, namely: Work, Expression, Manifestation, and Item. Ideally the MARC Clearinghouse would be develop and/or depend upon a community of groups who support each other in doing MARC transcoding to URIs. A web-based app could be built that helped institutions and projects make and provide quality control for URIs, then launch them into open stores for use by others. The community supporting this notional MARC Clearinghouse would share experiences as well as tools and/or tool development.

ADDITIONAL POTENTIAL PROJECTS

DOMAIN SPECIFIC PROJECTS

a. There are respiratory societies in the UK that are using Linked Data in their environments. A census of these is needed, maybe gained by working closely with JISC. An excellent horizon scan sponsored by JISC may be found here: http://linkeddata.jiscpress.org/ ; and another useful JISC site is: http://www.jisc.ac.uk/whatwedo/programmes/inf11/jiscexpo .

b. Involvement in any NLM activities.

c. There are scientists who are actively looking outside of their areas of expertise, and Linked Data helps to do the pivoting across subjects. A census of these is needed too.

d. Civil War 150 – Linked Data history project. Freebase is already involved. Push data to them so there can be a use case in the pipeline. See: http://www.history.com/interactives/civil-war-150#/home .

e. Nines – this is an existing use case; see http://www.nines.org/ .

f. Specific projects integrating existing ontologies and thesauri into LOD projects, with resulting URI creation and promulgation as LOD would be most useful. Willing and able "owners" of ontologies and thesauri need to be recruited and projects devised.

g. Arabic Union Catalog has been undertaken at the Bibliotheca Alexandrina with output as MARC records; transcoding those records could be undertaken.

h. There may be an opportunity for Tibetanists / Himalayan Studies to collaborate with David Germano at UVa and the Tibetan diaspora community. This needs investigation by an interested party.

i. ResearchSpace, a major Linked Data project of the British Museum, deserves attention and we need to investigate other LOD projects at museums and galleries. It is expected that including museums and similar cultural agencies in the desired prototype will make obvious the benefits of the unifying effects of a Linked Data discovery environment. See: http://www.researchspace.org/project-updates/museumsandthesemanticweb-britishmuseumstudyday .

j. MyExperiment.org in the UK is a pretty well advanced project. Getting into e-science, but they're Linked Data compliant. Would it be synergistic with our projects?

l. Bringing in the Getty thesauri would be a huge addition. Getty may have agreed to open their data, but this needs verification. See: http://www.getty.edu/research/tools/vocabularies/faq.html .

LINKED DATA CAPACITY BUILDING

Expanding the capability of the library community to publish, enhance and leverage Linked Data. Institutions are coming from different places, with different capacities. Recognize need to create workshops, tools, learning opportunities, or simply donate data for the creation of URIs.

READINGS AND REPORTS

In preparation for the Workshop a survey was performed by Jerry Persons with Mellon Funding through CLIR. That survey and this report are now public documents. And this report can best be understood in the light of the Persons Survey, which is being released simultaneously. See: http://www.clir.org/pubs/archives/linked-data-survey/part00_01_introduction.html

See the tab of the Persons Survey at CLIR for recitals of projects, completed and on-going, to date.: http://www.clir.org/pubs/archives/linked-data-survey/part10_projects.html

RELATED TOOLS

Tools listed below were identified in discussion as potential resources for projects going forward. In general, the group would like to see a central resource for identifying tools.

- The eXtensible Catalog (http://www.extensiblecatalog.org/) was originally funded by Mellon and is open source. It is interesting for its Metadata Management possibilities. The Metadata Services Toolkit enables the XC user interface to present FRBR-ized, faceted navigation across a range of library resources. The toolkit aggregates metadata from various silos, normalizes (cleans-up) metadata of varying levels of quality, and transforms MARC and DC metadata into a consistent format for use in the discovery layer. Their transformation of library metadata to RDF should be very high quality. A MARC Clearinghouse (Data Store) should be set up from the RDF derived from the iterative process outlined in Appendix A. The FRBR Group 1 entity relationships between resources should be included; those relationships are: Work; Expression; Manifestation; and Item.

- The Bibliothèque nationale de France released a first version of its "Linked Open Data" project: http://data.bnf.fr. The project includes simple Web pages about major French writers and works, applying FRBR principles. The HTML is fully open to the Web Example: http://data.bnf.fr/11910267/jean_de_la_fontaine/ For each page/concept, the RDF is available in RDF-XML, NT, N3:
    - http://data.bnf.fr/11928016/jules_verne/rdf.xml,
    - http://data.bnf.fr/11928016/jules_verne/rdf.nt,
    - http://data.bnf.fr/11928016/jules_verne/rdf.n3.

- The LUCERO Project (http://lucero-project.info/lb/2011/07/final-product-post-tabloid/) called TABLOID looks very good.
- http://consulting.talis.com/2011/09/putting-links-into-linked-data/
  The 'LOD Around The Clock' (LATC) Project, of which Talis Consulting is a part, is working to make it easier for dataset publishers to interlink their data with other datasets by developing a Linking Platform that will take care of the heavy lifting. See http://lod2.eu
  http://lists.w3.org/Archives/Public/public-lod/2011Oct/0021.html
  The LOD2 consortium [1] is happy to announce the first release of the LOD2 stack available at: http://stack.lod2.eu.
  The LOD2 stack is an integrated distribution of aligned tools which support the life-cycle of Linked Data from extraction, authoring over enrichment, interlinking, fusing to visualization. The stack comprises new and substantially extended tools from LOD2 members and 3rd parties. The LOD2 stack is organized as a Debian package repository making the tool stack easy to install on any Debian-based system (e.g. Ubuntu). A quick look at the stack and its components is available via the online demo at: http://demo.lod2.eu/lod2demo . For more thorough experimentation a virtual machine image (VMware or VirtualBox) with pre-installed LOD2 Stack can be downloaded from: http://stack.lod2.eu/VirtualMachines/

  Also see the accompanying survey under the TOOLS heading at
  http://www.clir.org/pubs/archives/linked-data-survey/part11_a_tools.html

CONCLUSION

"Of making many books there is no end, and much study is a weariness of the flesh."[6][7] Yet, this is one of the central ethical tenets of scholarship and teaching, one that professionals in the supporting disciplines of librarianship, museology, archival practice, academic publishing, and information management have tried to make sense of for centuries. The participants in this workshop have produced a number of insights and encouragement to addressing the hypothesis that an open Linked Data discovery

[6] Ecclesiastes 12:12 (English Standard Version 2001)
[7] For "books", please understand all products of modern scholarship including articles, maps & GIS reports, hypertexts, multi-media presentations, documentary movies, and more.

and navigation environment might save the time and effort of scholars and students in the pursuit of knowledge and information for their academic purposes. Given the proliferation of URIs, whether RDF triples or more, from numerous sources it seems plausible to attempt to model and then construct a discovery and navigation environment for research purposes based on the open stores of RDFs becoming available. To many of us, this seems a logical next step to the vision of the hypertext/media functions in a globally networked world of Vannevar Bush, Ted Nelson, and Douglas Englebart. It is highly significant to us as well that Tim Berners-Lee, responsible for the launch of the World Wide Web, has led this line of thought through his publications and presentations and those of his colleagues at the University of Southampton, Wendy Hall and Nigel Shadbolt. [8]

[8] Berners-Lee, Tim; James Hendler and Ora Lassila, *Scientific American*, May 2001, . "The Semantic Web".; and "Tim Berners-Lee on the next Web" February 2009 TED Conference, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html; "The Semantic Web", a talk at the Annenberg School at the University of Southern California by Professor Wendy Hall, http://www.youtube.com/watch?v=-XPc9d526lI ; Shadbolt, Nigel and Tim Berners-Lee, *Scientific American*, October 2008, pp 76-88.

APPENDICES

APPENDIX A: SAMPLE WORKFLOW FOR THE CREATION AND ITERATIVE RECONCILIATION OF RDF TRIPLES

1. Release early, release often

The deployment of Linked Data technologies has not been sufficiently widespread that problems are generally predictable; it is important to have sight of downstream issues at a stage when the investment in upstream processes is kept to a minimum.

The capabilities of the technologies are only beginning to emerge; the library professionals and their users need to see early outputs, so that they can feed back new ideas to the whole process.

2. Mint URIs

Choosing to mint a new URI as an identifier is usually a simple and quick decision, allowing the triplification process to continue at pace; trying to re-use existing URIs complicates the triplification process, and delays release.

Identifying appropriate URIs to re-use is error-prone, and can undermine the quality of the triples produced.

Using your own URI is simply saying what you want about your resources; this is less controversial than saying things about others' resources.

Where you use existing URIs, spend time reviewing them for accuracy.

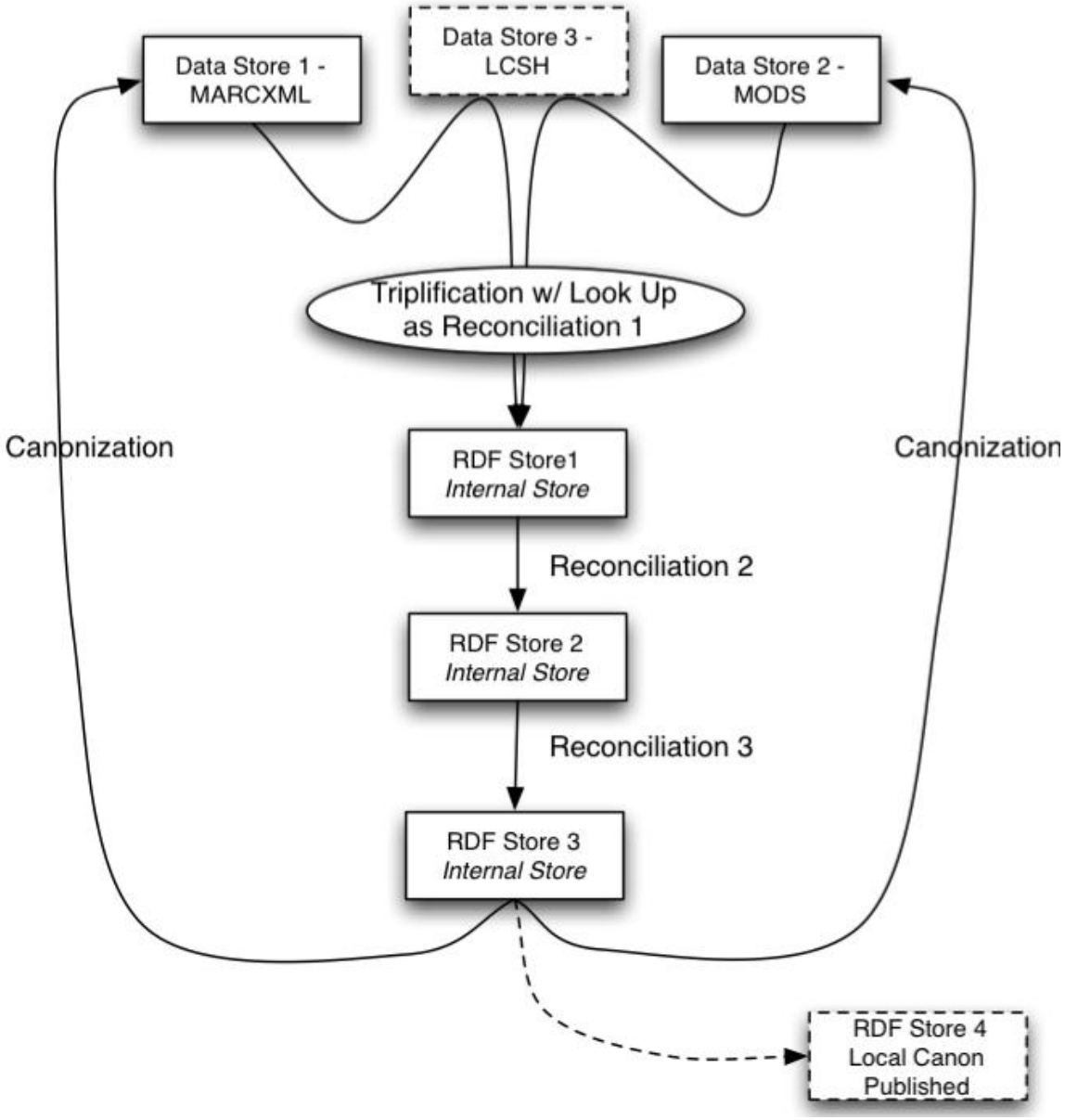
3. Leave linking to later

Linking is hard. Don't do the hardest thing first.

It needs lots of knowledge, some of which may improve as the process goes on, improving the linking in terms of false negatives and false positive.

Someone else may do it for you – or may even have already done it.

A Process:



1. The first stage is to translate the fundamental records in (MARC or whatever) into RDF. It is expected that as a result of this or other projects, existing tools can be deployed to do this. An ontology is required, but again, some standardization for library records is emerging.

   As part of this stage, URIs will be minted whenever there is doubt as to equivalence with external sources.

2. However, classifications such as LOC will clearly be used in the catalogues, and can safely be looked up to use "official" URIs, such as those provided at http://id.loc.gov/.

   This is safe and relatively cheap computationally.

3. Once this has happened, the RDF store that holds the data can be provided as an early release to appropriate partners. This enables early feedback on problems, and early development of visualization and services, identifying further problems and opportunities.

4. There now follow stages of data (or more accurately knowledge) enrichment, concerned with improving the co-reference information (reconciliation).

5. Machine-based algorithms are applied to identify co-reference (asserting skos:exactMatch or owl:sameAs or equivalents), where there is sufficient confidence in the result. These always work over the RDF store, as that is where the knowledge is held to inform them.

6. Further reconciliation can finally take place, where humans may be involved.

   This should always come as late in the process as possible:- it is foolish to have humans doing what can be achieved by machine, but more importantly, up until this stage, should the early stages change, any activity can be replayed easily. Once human effort is put in, it is harder to capture the process and replay it.

7. Apart from the cost, when a wide range of domains is involved, using humans is not as reliable as it is often thought to be, and so should be used with care. Systems that ask humans to verify or reject borderline matches, rather than add data de novo, are frequently the most productive.

8. Recording pairs of URIs that might have been thought equivalent, but have been found to be distinct, is very valuable.

9. The reconciliation stages might include: Lookup; Normalization; Simple Matching, Semantic Matching, By Hand.

10. As the reconciliation proceeds, the number of URIs that are found to have duplicates will increase, and it may prove useful reduce them. This process has been termed canonization.

11. This can be done by feeding the co-reference information back to the start of the process, and then essentially treating it as a Look Up, completely discarding the disregarded URI for the later stages.

12. At one of these stages, but hopefully as late as possible, URIs will start to be used by external systems that will then expect them to be maintained – essentially this is the publishing moment.

13. URIs that have been the subject of reconciliation can then no longer be discarded, although they can still be used for Look Up in the first stage.

Notes

1. Problems will arise in the quality of the source data. It may be that the catalogue identifiers have been re-used over the years, or that there are simply quite a lot of mistakes. In this situation, many more URIs than expected will need to be generated by algorithm from record fields, and so the reconciliation will be more extensive than expected.

2. The whole process will be replayed on a continuous basis, as more data arrives in the Data Stores. It is likely that the simplest way to do this is to do the recapture (with canonization). Since the reconciliation information is out with the stores, it will still apply to the newly recaptured RDF.

3. A triple with a string in the object position should only be used if the predicate can sensibly be made a subclass of rdfs:label. For example, if I assert that <URIa has-author “George Orwell”>, I am unable to assert that this author of URIa is the same George Orwell as URIb. The whole point of Linked Data is that everything has a URI. I should have asserted something more like <URIa has-author URIc> and <URIc rdfs:label “George Orwell”>.

4. Being able to explore and visualize the data (for the technologists and library professionals, but not necessarily end-users) is an early requirement, as the process needs to be informed by what is emerging in the RDF store.

5. Free text search is not a strength of most RDF stores, and so the RDF store may need to work in tandem with something like SOLR.

APPENDIX B: LINKED AND OPEN DATA IN RELATION TO CULTURAL HERITAGE VENUES

For additional information, see the Extant metadata, Sources of identifiers and links, and Projects … in the Literature Survey that accompanies this report. The *open* and *linked* columns in the table refer to data/projects for which the intent is to produce open (*i.e.* CC0) data, and/or produce that data in some form of Linked Data. CKAN refers to Comprehensive Knowledge Archive Network and its Data Hub.

Europe

	open	linked
CENL (Conference of European National Libraries) representing 46 libraries voted "to support open licensing of their data" on 28 September 2011	x	

Library Data

	open	linked
British Library	x	x
Cambridge University Library (CKAN)	x	
CERN bibliographic data (CKAN)	x	x
data.bnf.fr (Bibliothèque nationale de France)	x	x
Deutsche Nationalbibliothek		x

hbz Union Catalog [Germany] (CKAN)	x	
National Library of Hungary (NSZL)		x
Open Library (Internet Archive) (CKAN)	x	x
Swedish Union Catalog (LIBRIS)		x
Talis MARC records (5.5 M) (CKAN)	x	
University of Michigan original cataloging (CKAN)	x	
Universitäts- und Stadtbibliothek Köln (CKAN)		x

Journals

	open	linked
arXiv	x	
DOIs … as linked data	?	x
HighWire Press	x	

Authority Files

	open	linked

Bibliothèque nationale de France: RAMEAU (subject authorities)	x	x
Deutsche Nationalbibliothek: name & subject authority files		x
JISC Names Project	x	
Library of Congress: name & subject authority files		x
New York Times subject descriptors	x	x
OpenCyc	x	
UMBEL (Upper Mapping and Binding Exchange Layer)	x	x
Virtual International Authority File (OCLC)	?	x
VIVO (Cornell and elsewhere)	x	x

Cultural heritage, research data & Linked Data initiatives

	open	linked
ANDS (Australian National Data Service)	mixed	mixed
BBC	x	x

Chronicling America: historic American newspapers	x	x
CKAN cultural heritage groups (archeology, art, economics, history, linguistics)	mixed	mixed
DataCite	x	
eGovernment initiatives	mixed	mixed
Europeana	x	x
Freebase		x
Geonames	x	x
LOACH (JISC: Linked Open Copac Archives Hub) … EADs as Linked Data	x	x
National Archives of Great Britian	?	x
OAIster (OCLC, was at University of Michigan)	?	
OKF (Open Knowledge Foundation)	x	NA
RDTF (Resource Discovery Taskforce) UK research library metadata initiative	x	?

ResearchSpace (support for cultural-heritage research)	x	x
<SameAs>	x	x
Talis: linked-data platform and Kasabi data market	mixed	x

APPENDIX C.: PARTICIPANTS

Charles Henry chenry@clir.org [unable to attend the Workshop, active partner in building the agenda and hosting]

Chuck is the President of the Council on Library and Information Resources. A brief, but outdated, bio for Chuck can be found at http://www.clir.org/news/pressrelease/06henrypr.html .

Michael A. Keller Michael.keller@stanford.eduMike Keller is the University Librarian at Stanford and wears a few other hats there. A brief bio can be read at http://highwire.stanford.edu/!mkeller .

Mike and Jerry Persons have been gnawing for three plus decades at the problem of discovery for scholarly purposes in an environment of: increasing numbers of silos of content & metadata; increasing complexity of search; the pervasive, yet erroneous belief that Google indexes all; and despite it all the realization in the research communities of the potential of interdisciplinary research. For the past several years Jerry & Mike have focused upon Semantic Web and Linked Data possibilities as possible means to start afresh and to create a much richer, more extensive, and, for the user, a more simple approach to discovery.

Jerry Persons jpersons@stanford.eduJerry is the Chief Information Architect Emeritus of Stanford University Libraries, where he had a distinguished career in that role, as head of the Library Systems Office, and as the Head of the Music Library for over 30 years.

Hugh Glaser hugh.glaser@seme4.com

Hugh Glaser has more than 30 years experience in Computer Science. His research work has most recently been as a Reader in the School of Electronics & Computer Science at the University of Southampton, UK.

His earlier research was in the fundamentals of Distributed Systems and Programming Languages, but since the Semantic Web activity began he has moved his focus to the technologies required to deliver the vision. As part of this he has enthusiastically embraced the Linked Data initiative.

In addition to the general work and consultancy he is responsible for two significant practical activities in the Web of Data: a) sameas.org , which helps to establish linkage

between datasets; b) rkbexplorer.com, which is a Linked Data application that gives a unified view of some fixed datasets plus data from the general Web of Data

Noha Adly Noha.Adly@bibalex.org

Dr. Adly is Deputy Head of the ICT Sector, Bibliotheca Alexandrina and a Professor of Computers and Systems Engineering, Alexandria University, Egypt.

I have joined the Bibliotheca Alexandria since 1997 as a Consultant for the design and installation of its network and information system. Since 2001, I have been sharing the responsibility of the BA's overall ICT strategy, architecting its data policies, including its institutional repository and digital library systems. Work at the ICT Sector is being driven by the Library's vision of building a universal digital library. This is being manifested through a variety of projects and research endeavors which aim at access to knowledge to all using state-of-the-art technologies. In this context, the ICT Sector's work comprises the creation of searchable documentary digital archives and repositories which encompass cultural preservation, in addition to science-oriented endeavors which serve researchers and scientists.

We have built the BA's Digital Assets Repository (DAR) system in-house based on open source tools since 2004. We have been continuously releasing upgraded versions of the system for accommodating the ever growing diverse collections of the BA ever since. DAR's core architecture involves grouping the different application silos into integral sets, applying unique identifiers to objects and heavily relying on triples and RDF in relating digital objects and their components. Hence, Linked Data fall right into the scope of our digital library technological philosophy. Moreover, our expertise is significantly articulated in the digitization of Arabic content, where we have partnered with several institutions in that context, like the Institut du Monde Arabe (IMA), Wellcome Trust, World Digital Library (WDL), to name a few. I believe that such content would represent quality raw material for Linked Data.

Magdy Nagy Magdy.Nagy@bibalex.org

Dr. Nagy is a Professor in the Computer Science department, Faculty of Engineering, Alexandria University. He obtained his Ph.D. from the University of Karlsruhe, in 1974, where he served as Lecturer for two years and as a Consultant to its Computer Center from 1974-1990. During this period he also served as Consultant to many companies in Germany such as Dr. Otker, Bayer, SYDAT AG, and BEC.

On the national level he was a Consultant to many projects under the umbrella of either the University of Alexandria or the Faculty of Engineering for designing and/or implementing automation projects for governmental authorities or public sector companies, such as the Ministry of Interior, the Health Insurance Organization (HIO), the Social Insurance Organization (SIO), and the Customs Authorities.

Since 1995, Dr. Nagy has served as Consultant to Bibliotheca Alexandrina. Among his activities are the design and installation of Bibliotheca Alexandrina's network and its information system as well as the design and implementation of the library information system, namely a trilingual information system that offers full library automation. He is currently serving as the Head of the Information and Communication Technology (ICT) Sector at Bibliotheca Alexandrina.

Dr. Nagy is a member of the ACM and the IEEE Computer Society as well as several other scientific organizations. His main research interests are in operating systems and database systems. He is author/co-author of more than 80 papers.

Eero Hyvönen eahyvonen@cc.hut.fi

Home page: http://www.seco.tkk.fi/u/eahyvone/

Eetu Mäkelä eetu.makela@aalto.fi

D.Sc. Eetu Mäkelä has been working on linked cultural heritage data for eight years now. Particularly, his interests have been focused on discovering what new functionalities Linked Data can give to human end-users, and how these can be realized. This has necessitated a broad view, from understanding user needs and interface design to dealing with the thorny issues of integrating massively heterogeneous data on a quality level sufficient for new possibilities to emerge.

As a concrete example, Eetu is the chief architect behind CultureSampo (http://www.kulttuurisampo.fi/?lang=en). This portal gathers together some 600 000 items of heterogeneous cultural heritage content from some thirty different institutions in about twenty different original schemas into a unified whole. Among the data gathered are museum items, historical news paper articles, poems, paintings, videos and even semantically annotated skills. Using the data selection, visualization and exploration functionalities of the portal a user can then use the unified data repository to discover for example:

* how imports from Japan to Finland have changed in the 20th century

* what were the most popular themes addressed in different forms of culture in Finland in the year 2007

* how beard fashions in Finland changed in the late 19th century and * what is the place of the mythical character Väinämöinen among all Finnish culture.

Joan Smith jsmit52@emory.edu

Formerly Chief Technology Strategist for Emory University Libraries and affiliated faculty in Computer Science. On the Libraries side, I managed both strategic planning and operational implementations of the libraries' technologies from digital scholarship sites to the OPAC and discovery Tools from 2008--2011. With the recent hire of an operations manager, I have migrated into a primarily strategic and R&D role for the libraries. As CS faculty, I designed and teach the graduate Software Engineering course (every Fall) as well as a "special topics" graduate course each Spring. One of my key activities in this dual role has been to integrate student course work with the Library's technology needs. I have an additional role as PI on a grant from the Mellon Foundation to develop a Digital Scholarship Commons ("DiSC"), which focuses primarily on digital humanities projects at Emory.

My involvement with Linked Data has been limited to a few recent R&D projects at Emory: A local implementation of VIVO; an Emory-branded prototype of Harvard's RNS Profiles software; and a quick-stab attempt to create a FOAF based on our OPAC. We've begun to add RDF-based features to

our Visual Shelf Browser project, but we're not really on the road yet (I'd say we're still tying our shoes). We are migrating to a new ILS and are working on substantially revising our metadata practices to encourage, incorporate, & take advantage of, linked open data. In short, there has been lots of interest but not a lot of action until very recently.

Professionally my focus has been (a) software process/methodologies and (b)preservation, but I am now becoming closely involved with our own Linked Data project and the development of an Open Access repository. My latest twitter account is "@joansm1th" and I occasionally blog at the R&D team's new Blogger site: stacks4libs. I'm fairly active on LinkedIn but have avoided my FaceBook account for years and gave up on MySpace ages ago. Prior to Emory, I spent many years as a software engineer/director of engineering at various technology firms. More info about me is at http://www.joanasmith.com/.

Stefano Mazzocchi stefanom@google.com

Stefano currently works as a Software Engineer at Google.

Previously, he worked as an Application Catalyst at Metaweb Technologies Inc. tasked to help enabling a development ecosystem around Freebase as a platform. Metaweb was acquired by Google in 2010.

Before that, he was a research scientist at MIT working on the SIMILE Project for the Digital Library Research Group of the MIT Libraries.

He is also known for his open source activities within the Apache Software Foundation (ASF) of which he's been a member since 1999 and a director between 2003 and 2005. There, he's mostly known for having started the Apache Cocoon project and, before that, for having being a release manager for Apache JServ (a servlet container now retired, precursor of Apache Tomcat), but some of his code can be found in several open source projects.

He has also participated in several expert groups within the Java Community Process, such as the Servlet API, the Java XML API and the Java Content Repository API.

His research interests include data integration, data mining and data visualization, virtual communities dynamics, software usability, user interface design and software engineering.

Reilly Hayes rlyeh@google.com [unable to participate in the Workshop]

Reilly Hayes was v.p. for data at Metaweb from the quantitative data driven world of program and algorithmic trading. Reilly created innovative trading products at Schwab, B of A, and Merrill Lynch. While comfortable with the enterprise, he is the veteran of four startups, including his own trading technology firm. Prior to working in financial technology, Reilly worked as a developer on an early mini-computer RDBMS, a mini-computer OS from DEC, and a software startup from the launch of the PC era. When Google acquired Metaweb, Reilly moved with the company into Google.

Jamie Taylor jamietaylor@google.com

When the world's information is available as structured data interesting things start to happen. Open Data, as a resource, flows between providers and those who can increase its value, forming markets. Data providers can reap value through the work provided by external data consumers. Data consumers obtain value through access to new data sets which can fuel value added services. My long term interest is in providing data infrastructure and services which facilitates this type of market ecosystem. To that end, I have been involved in developing systems that create and expose structured data both within large enterprise systems and in large scale public systems. I help launch one of

the first Bay Area ISPs providing dedicated ISDN service to downtown San Francisco and have been an evangelist for Freebase and semantic data in general. Over the years I have provided technical consulting for companies like CSC, CapGemini/Ernst and Young and co-authoring the O'Reilly book 'Programming the Semantic Web.' I currently work for Google and hold a Ph.D. from Harvard University in Behavioral Economics.

Akihiko Takano aki@nii.av.jp

Prof. Akihiko Takano is Professor and Director for Research Center for Informatics of Association at the National Institute of Informatics in Japan. Prior to joining NII in 2001, he had worked at research laboratories of Hitachi, Ltd. for almost 20 years. He holds a B.A. in Mathematics and a Ph.D. in Computer Science, both from the University of Tokyo. Since 2002, he is also Professor at Department of Computer Science, the University of Tokyo.

Rachel Frick RFrick@clir.org

Rachel Frick is the Director of the Digital Library Federation Program at the Council on Library and Information Resources(CLIR/DLF). Prior, to CLIR, Ms. Frick was the senior program officer for the National Leadership Grants for Libraries, at the Institute for Museum and Library Services (IMLS) . Ms. Frick's library experiences range from being the head of bibliographic access and digital services at the University of Richmond to a regional sales manager for the Faxon Company, with a variety of library positions in between. She holds an MSLS degree from the University of North Carolina at Chapel Hill and a BA in English literature from Guilford College.

Kevin Ford kefo@loc.gov

Kevin Ford works in the Network Development and MARC Standards Office (NDMSO) at the Library of Congress where he is the current project manager for the Library of Congress's Linked Open Data service, ID.LOC.GOV, which publishes LC owned or managed authority and vocabulary data as Linked Data. ID.LOC.GOV includes LC Subject Headings; Thesaurus of Graphic Materials; MARC Lists for Relators, Geographic Areas, Countries, and Languages; ISO 639 Language codes, parts 1, 2, and 5; and PREMIS vocabularies. He is responsible for data conversion of ID data to MADS/RDF and SKOS. In addition to all technical development for LC's Linked Open Data service, Kevin spends a significant amount of time modeling traditional library authority and vocabulary data in RDF for publication at ID and consulting within the Library on other vocabulary-related issues.

Kevin also participates in standards development within NDMSO. He was part of the development team for MADS/RDF. Recently, Kevin has contributed to the development of a PREMIS ontology expressed in OWL, specifically because of its reliance on data values published as part of LC's ID.LOC.GOV service. He has also contributed to early drafts of MODS in RDF. Kevin is participating member of the W3C's Library Linked Data Incubator Group

Adam Soroka ajs6f@eservices.virginia.edu

Adam Soroka is the Senior Engineer for the Online Library Environment group at the University of Virginia Library, with particular responsibility for digital object repository architecture and workflows. He has taken up that role this year after completing the first round of development of Neatline, an NEH and a project funded by the Library of Congress examining the intersection of bibliographical modeling and geo-temporal visualization. He is a member of communities around several technologies with interest for Semantic Web systems, particularly the Fedora Commons repository framework.

In connection with Semantic Web technologies, his research interests include the intersection of Linked Data with geospatial Web services, the use of RDF for modeling complex bibliographical systems, translations between RDF and traditional structural metadata markup like EAD or METS, and the use of markup editing tools with such translations to provide community-specific applications over RDF stores.

Bill Dueber dueberb@umich.edu

Bill Dueber (University of Michigan) is a Systems Librarian working primarily as developer of front- and back-end systems that comprise metadata catalogs for the University of Michigan and the HathiTrust (at catalog.hathitrust.org).

I will be a co-designer and technical architect as well as lead developer of infrastructure and applications that support linked access to metadata (bibliographic, holdings, and access rights)

Dave Price Dave.Price@bodleian.ox.ac.uk [unable to participate in the Workshop]

Head of the Systems and eResearch Service of the Bodleian Libraries, which has general responsibility for IT, the Libraries' business applications and digital library systems. The Bodleian's core architecture for the storage and preservation of digital objects is based on an RDF structured object store. As a result, Linked Data and semantic web technologies form a core part of our technology stack and digital library

strategy. Particular services include our institutional repository (ORA), and associated research information repository (the BRII project) and nascent research data repository (DataBank). We also host the University's ontology and vocabulary store (vocab.ox.ac.uk). We are involved in a number of projects which are constructing semantic knowledge models based around library resources, including:

- Cultures of Knowledge (www.history.ox.ac.uk/cofk) researching the 17th C republic of letters, a five year Mellon funded project with multiple European collaborators
- IMPAcT, re-using the CoK object model for 13th-16thC Persian manuscripts
- Fihrist (www.fihrist.org.uk), a union catalogue of Islamic manuscripts (based on the joint Oxford and Cambridge Islamic Manuscripts Catalogue Online project)
- Genizah, reusing the Fihrist model for Hebrew manuscript fragments
- DMSTech, collaboration with Stanford and others to develop a standard for describing the visual representation of manuscripts, including alternative binding sequences, foldouts and fragment re-assembly, and software to render that visualization.
- Medieval Libraries of Great Britain based on the print work of the same title, which reconstructs great medieval collections based on extant manuscripts and catalogues.
- Bodleian Incunable Catalogue, which will be producing an enriched online version of the original print work.

David Rosenthal dshr@stanford.edu

David Rosenthal is the Chief Scientist of the LOCKSS program at the Stanford Libraries, which provides libraries with tools to preserve web published materials (ejournals, books, blogs, web sites, archival materials, etc) for the long term. Long term preservation will be an important aspect of the proposed Linked Data environment.

David was an early employee and Distinguished Engineer at Sun Microsystems, and employee #4 at Nvidia before starting the LOCKSS program at the Stanford Library in 1998. He has worked on graphics
software and hardware, file systems, middleware, and system and network administration.

Ed Summers edsu@loc.gov

I am a software developer working at the Library of Congress in a digital preservation unit. I focus on providing access to digital materials in projects such as the National Digital Newspaper Project (NDNP), the National Digital Information Infrastructure Preservation Program (NDIIPP) [2], and LC's internal Content Transfer Services platform. It's my firm belief that digital preservation is a function of access to digital content; and that the Web is the ideal delivery platform for this content--for both people and machine agents.

I've actually used the Linked Data pattern in several projects:

- NDNP's public access webapp Chronicling America [3], which uses Linked Data (DCTERMS, BIBO, OAI-ORE) to provide access to the metadata and bit streams associated with 4 million newspaper pages, and their associated issues, titles.
- LC's Authorities and Vocabularies Service [4] which made the Library of Congress Subject Headings available as Linked Data (SKOS, DCTERMS)

I was a member of the Semantic Web Deployment Working Group at the W3C [5] that standardized SKOS [6] and RDFs [7] and am currently participating in the Library Linked Data Incubator Group [] at the W3C. Despite my interest in Linked Data, I'm not religious about RDF, and think that other metadata practices (Microformat, HTML5, Atom, JSON and XML) have their strengths. I am also a big fan of REST, which has allowed the Web to grow into the wonderfully rich, global information space it is today.

Jim Nisbet niz@stanford.edu

Jim Nisbet's role at HighWire Press, a division of the Stanford University Libraries, was to help HighWire make optimal technical decisions and help set technical directions and plans for the future. This focus includes semantic analysis of HighWire hosted scholarly content. Prior to HighWire, Jim has been involved with seed funding and technical due-diligence of startups and spent two years with Semio Corporation working on content classification software solutions. Since the Workshop, Jim has reverted to a leadership position in a Silicon Valley start-up company.

Jim was the Chief Technology Officer of RSA's Data Security Group. He was a founder and Chief Technology Officer of two successful companies: Tablus and DataTools. Tablus was a data security company acquired by RSA Security and DataTools was a database tools company and acquired by BMC Software.

Lars Svensson l.svensson@dnb.de

I'm an IT manager in the German National Library (DNB) dealing mainly with knowledge organization and persistent identifiers. I've been into RDF and the Semantic Web since about 2004, mainly looking at how
libraries can publish their authority and bibliographic data as Linked Data and which role persistent identifiers can play when doing that. I am well aware of the issues with Linked Data vs. Linked Open Data, which many libraries (including my own) hesitate to use an open license (and the reasons for not doing so), as well as the potential a non-restrictive licensing policy might have for organizations outside of the cultural heritage sector.

For technical matters: I've worked as a software developer (mainly Java) in the library environment, but I don't have explicit knowledge or experience with producing or consuming RDF data.

Phil Schreur pschreur@stanford.edu

I am currently the Head of the Metadata Department for the Stanford University Libraries. In this capacity, I am responsible for the creation of descriptive metadata for the traditional Stanford collections and an increasingly large number of digital resources. In order for Linked Data to be applied most accurately and efficiently, the links must be created by automated means. Library metadata, with its controlled access points, is an ideal place to begin. Many resources, however, lack these controlled terms. I am most interested in exploring automated and semi-automated means of assigning these terms to the vast array of resources not controlled by the Library. My work as the Knowledge System Developer for HighWire Press has shown that the assignment of controlled terms for concepts can be done through the semantic analysis of text on a massive scale. Through semantic analysis and the use of international authority files such as VIAF, quick and accurate links can be made between disparate resources.

[Since the Workshop, Phil has been appointed the chief organizer of SULAIR's Linked Data Projects.]

Richard Boulderstone Richard.Boulderstone@bl.uk

We (British Library) are providing free sample RDF formatted data from the British National Bibliography (BNB) http://www.bl.uk/bibliographic/natbib.html on our website at: http://www.bl.uk/bibliographic/datasamples.html. This is a trial to gauge

response from the community - if favorable we will provide the entire BNB through this route.

Richard Webber rwebber@stanford.edu

Richard Webber is Associate Director for Enterprise Systems and Programming at Stanford University Libraries. In this role he has responsibility for driving the overall strategy for developing, delivering and supporting enterprise level applications and services for the libraries. Richard is directly responsible for the Library Management System (SirsiDynix Symphony) including the metadata associated with over 6.5M titles that the libraries hold. He is also responsible for Stanford's course management/virtual learning environment (based on the open source Sakai Project) and has a team of developers, QA engineers and administrators that develop, test, deploy and maintain the application. This year, Richard is leading the creation of SUL's next generation enterprise systems platform based on VMWare and Oracle RAC, that will serve as the hosting environment for the majority of the mission critical applications and services that SUL supports.

Richard came to Stanford in November 2010 with over 20 years of industry experience at both smaller and larger software companies including Hewlett-Packard and Intuit. In these roles he has led the development of both self-hosted and SaaS based enterprise applications. Richard has managed software development, QA, release engineering and user experience and focuses on the end to end process of taking an idea to a supportable, reliable application.

Richard will be very involved in the complete lifecycle of the Linked Data project with special focus on the development process and how to take it into production and support it at scale.

Romain Wenz romain.wenz@bnf.fr

Since July 2009: Curator at the French national Library (BnF), working as a metadata expert at the Bibliographical and Digital Technology Information Department (IBN). This department is in charge of the metadata (both for books and for digital content). We work on the standards, production, and development of the authority files.
As the head of the "data.bnf.fr" product, I specifically work on building this "pivot" site, with a team from the IT department and developers. This is basically writing the specifications, testing, and working on it with the team. But it also includes communication, and interaction with similar projects.

I have a traditional Librarian background: I graduated from the Ecole nationale des Chartes (ENC, degree in Archives management, thesis in medieval history), from the French national Library school (ENSSIB, degree of state chief librarian), and from Paris Panthéon-Sorbonne University (Master's degree in history).
Among other experiences: Archival descriptions for the French national Archives (AN, Paris, 2004 and 2005), Curation of an antique weapon collection (Clermont-Ferrand, 2006), Cataloguing of Jean-Martin Charcot's collection of rare books (UPMC, Paris, 2007), general tasks in a public library (Cité des sciences, Paris, 2008), work on the TEL application profile (The European Library, The Hague, 2009).

Sigfrid Lundberg slu@kb.dk

Born 1956, I became Ph.D. in theoretical ecology 1985 and full time software developer/Internet programmer in 1995 at Lund University Libraries, Sweden. Specialized early on web harvesting, text retrieval and encoding and metadata. Active within DCMI 1996 to 2001. Became trans-national commuter 2005 when I started to work at the Royal Library, Copenhagen. My spare time is spent on family, music and photography.

I develop of software in Java and UNIX/Linux environments and prolific developer in (for example) Java, XSLT, Perl, Shell and SQL. Experienced at processing data, and metadata, for search and navigation encoded in TEI, METS, MODS etc. Have used REST based web services for years and a strong proponent for COOL URIs for the web of data. Most recent publication: RFC6120

Stephen Abrams Stephen.Abrams@ucop.edu

Stephen Abrams is the associate director of the University of California Curation Center (UC3) at the California Digital Library (CDL), with responsibilities for strategic planning, innovation, and operation of the center's services, systems, and projects. He designed for the center's micro-services-based Merritt curation repository, incorporating an OAI-ORE-based data model and a central Linked Data metadata catalog. Mr. Abrams is leading the Uniform Digital Format Registry (UDFR) project to create a community-supported semantic registry of format representation information useful for curation and preservation
purposes.

Tom Cramer tcramer@stanford.edu

Tom Cramer is the Chief Technology Strategist and Associate Director of Digital Library Systems and Services for the Stanford University Libraries. In this role, he oversees the technical development and delivery of Stanford's digital library activities, including the digitization, next generation catalog and discovery services, digital preservation, digital repository and digital asset management services.

With regard to Linked Data, Tom is exploring their use in three distinct spheres: leveraging existing open Linked Data in SearchWorks (Stanford's next generation catalog) to augment discovery services; relating digital objects and their components in a repository context for asset management; and application of the Open Annotation Collaboration data models to digitized medieval manuscripts, enabling cross-repository discovery, use and annotation of these materials.

APPENDIX D. WORKSHOP AGENDA SUMMARY & OVERVIEW

The workshop consisted of three major segments:

- Introductions and short talks took the first day, and focused on providing background on each participant's experience with and understanding of Linked Data. This segment established a baseline for discussion, and allowed participants to define the agenda going forward.
- Small group collaborative sessions were the heart of the meeting, and took three days. These discussions were intended to identify challenges and opportunities, and to define the business case for Linked Data
- The last day of the workshop was devoted to full group discussions to refine and codify the ideas brought forth in the small group sessions, prioritize issues and concerns, and outlining projects and partnerships

The sections below highlight the general outline of each day, pulling out key discussion topics and items of interest. Significant work products are called out, but are detailed in the next section, Workshop Products.

DAY ONE – MONDAY, JUNE 27TH

Introductions

The program opened with each individual introducing him or her self.

Agenda Setting

The group reviewed objectives, and discussed an outline for the activities for the week. It was agreed that the agenda would be flexible, and responsive to the ideas and concerns raised in each day's programs.

Groups

Four workgroups were established for the discussion portion of the program. All small group work referenced below was performed in these groupings

Group 1

- Leader: Hugh Glaser
- Tom Cramer

- Noha Adly
- Adam Soroka
- Rachel Frick
- Jamie Taylor

Group 2

- Leader: Jim Nisbet
- Richard Boulderstone
- Stephen Abrams
- Lars Svensson
- Reilly Hayes
- Eero Hyvönen
- Jerry Persons

Group 3

- Leader: Richard Webber
- Sigfrid Lundberg
- Romain Wenz
- David Rosenthal
- Kevin Ford
- Akihiko Takano

Group 4

- Leader: Ed Summers
- Magdy Nagi
- Phil Schreur
- Bill Dueber
- Joan Smith
- Eetu Mäkelä
- Mike Keller

Introductory Talks

Each participant was asked to give a very brief presentation about their involvement with Linked Data and concerns related to Linked Data.

DAY TWO – TUESDAY, JUNE 28TH

The group broke into working groups and worked through the following questions:

- What are the challenges of Linked Data, and what are the opportunities?
- What is the scope of the issue that this group should address, and what is the business case for Linked Data?

The working groups came together after looking at each question, to compare notes.

Key Discussion Points

Prototyping and feedback are important. Build something that people can complain about!

Even without an economic model, a number of us have provided data. Where are the applications that consume that, and what can we do to facilitate that?

There are a bunch of vendors out there who just create junk triples. What differentiates is maintaining quality. Provenance/trust, the historical record, and the correction cycle all feed into this.

There is a tension between human curation and machine generation. High quality data appears to require the former; doing anything at scale requires the latter.

Entity resolution is hard. It's the problem of publishing Linked Data without linking to anything.

For many cultural heritage organizations, the business case is not about dollars. We need to outline the goals and objectives of the organization, and demonstrate the that the project advances them.

DAY THREE – WEDNESDAY, JUNE 29TH

Agenda Setting

Joining the Linked Data World

- What challenges/obstacles did you see before?
- What new challenges/obstacles do you see now?
- What opportunities do you see?
- What scope does this group define?
- Linkage among specific projects?

Goals, business & use cases…
assuming common specifications, requirements, & protocols and assuming collaborators:

- Speaking (anonymously) as though you are the leader of your organization, how would a Linked Data project affect your organization's means of achieving its goals?
- Speaking as above, what are the business cases you need to propel enthusiasm & funding for your Linked Data project?

Speaking as above, what are the use (actor, action, benefit) cases that would propel enthusiasm & funding for your Linked Data project?

Semantic web linking standards: "In a sea of RDF triples, no developer is an island".

Challenges/opportunities:

Considering goal (specifications, requirements, basic design) for open LD/rdf stores

1. Transcode (generate RDF3s, URIs) MARC records?
2. Transcode article metadata records (many unknown formats)?
3. Transcode other available metadata?

What demand might there be for LD environments per discipline or generic agency, e.g. bio-medical, geo-spatial, clinical trial results, linkage to art galleries, museums, other cultural agencies?

Integration with known services & programs (e.g. Google, Seme4, Finnish LD programs, etc.)

Coordination of effort among institutions & companies

LDW known Issues:

- Quality Control
- reconciliation
- Licensing
- data curation
- scalability
- attribution/origin/authority
- staff training
- relationship to e-scholarship (esp. e-science) & e-learning
- Quality Control of unfamiliar languages of metadata or info objects
- Use of library authority files – names, subjects, etc.
- Use of ontologies

[The group decided to review the top three points first.]

Key Discussion Points

The interplay between Linked Data and MARC was one of the more heated discussion topics for the day. While Linked Data provides a pathway away from MARC, the transition will not be instantaneous. There will need to be a workflow and an audit trail built out.

PRIORITIZING THE KNOWN ISSUES

Richard Webber led a process by which the group came together as a team to identify priorities. The priorities will not constrain work, but will be able to guide the thinking of the teams, and bring some consensus. Results of the process are found below.

DAY FOUR – THURSDAY, JUNE 30, 2011

Mike Keller opened the day by describing a vision. There is a need to have a framework that can be put in the hands of people in four groups:

- Haven't created triples, and starting from scratch
- Have triples, but haven't got an application

- Have a domain or application, and want to make it more useful/effective
- Have some data, and want to make it more useful

Plan to work through the prioritized Known Issues list, and coordinate between the list and the categories. Groups can either work in order through the list, or pick from the list based on subject knowledge.

Assignment:

Confront the "operational matters" and begin to put some definitions against them so that over the next several hours we can assemble the list in a framework against the three categories.

Group 1 came up with a useful workflow for deploying Linked Data that is highlighted below.

DAY FIVE – FRIDAY, JULY 1, 2011

The team watched two presentations, one from Hugh Glaser, and one from Eero Hyvönen. Hugh's presentation is summarized in the flow chart and accompanying notes above. Eero's was based on more than a decade of building an operating numerous Linked Data sites n the ONKI "Finninsh Ontology Service". His findings and advice is encapsulated here and is significant for anyone embarking on a Linked Data project:

- **futures**: data ←→ prototypes (aka use cases)
    - BL + HighWire + [library data] + [museum data] + authority files + places
    - harvesting the data sets
    - "RDFIzing" … includes replacing literal strings w/ URIs **throughout** data
    - vocabulary alignment (keywords, authorities, places, events, …)

- metadata schema alignment
- data validation
- reconciliation

-- sameness/matching

-- de-duplication

-- all at scale plus at very high precision

- **tools**: [in ONKI toolbox]
  - RDFizing for MARC21
  - text annotation … [unclear to me either purpose or tool]
  - metadata editing via SAHA
    -- distributed RDF metadata editing environment
    -- http://saha.googlecode.com/files/saha_technical_report_2011_05_18.pdf
- **summary:**
  - Eero & Eetu have nearly a decade of experience working with disparate LAM organizations and their metadata, debugging the processes of filtering the vagaries into useful pools of metadata and working with disparate LAM cultures
  - same for mucking about with all manner of linked and semantic data approaches, tools, vocabularies, ontologies, etc.
  - Eero & Eetu have valuable experience, one of many to be consulted as planning for projects and futures go forward