Stanford University

LIBRARIES & ACADEMIC INFORMATION RESOURCES

Stanford Linked Data Workshop Technology Plan

30 December 2011

Table of Contents

Introduction

This is a plan for a multi-national, multi-institutional discovery environment built on Linked Open Data principles. If instantiated at several institutions, will demonstrate to end users the value of the Linked Data approach to recording machine operable facts about the products of teaching, learning, and research. The most noteworthy advantage of the Linked Open Data approach is that it allows the recorded facts , in turn, to become the basis for new discovery environments. This model includes the basic functions of generating, harvesting, and iteratively reconciling URIs as well as consumption of Linked Data. Consumption involves adapting or building one or more "killer apps" (user interfaces harvesting and displaying relationships among the products, services, and staff of research institutions). The model also provides guidance for assembling and/or adapting tools supporting the necessary steps in workflows emitting RDF triples and related URIs to open stores for use by any discovery service. The resulting discovery environments will demonstrate the dramatic change that is possible in the academic information resource discovery environment when organizations move beyond closed and rule-bound metadata creation and utilization. We believe that these closed operations are limiting and detrimental to the academic or research processes they are meant to support. This model also postulates dramatic changes to the creation, adoption, editing, and maintenance of metadata records for bibliographic holdings as well as scholarly information resources licensed for use in research institutions; there are indicators of revisions in this Plan to the classic cataloging services (and other operations of research libraries' technical services divisions – acquisitions, serials control, inventory control and circulation, auditable financial transactions, etc.) as well as metadata generation and distribution by scholarly journal publishers and their service providers.

This model was developed in conjunction with the Linked Data Workshop conducted at Stanford University 27 June through 1 July 2011, with support from the Andrew W. Mellon Foundation's Scholarly Communications Program, the Council on Library and Information Resources, and the Stanford University Libraries.[1] In addition to the Workshop, a Literature Survey was produced to inform first the Workshop participants and then the research library community of "the practical aspects of understanding and applying Linked Data practices and technologies to the metadata and content of libraries, museums, and archives."[2]

We postulate an institutional base for this model, but expect that many institutions would adopt and implement it. The model does not require elaborate coordination mechanisms once the basic data model is ingested and adopted by schema.org. Once adopted by schema.org, we expect the model to evolve as RDF triples and URIs, as well as variant data models, are proposed. Schema.org's role, one it plays already, is to constantly evolve a universal data model based on submissions of new versions pertinent to genres, formats, and needs of its contributors. This model does not require implementation by numerous institutions; we believe that implementation by a relatively small number of research

[1] A report of the Workshop's output and processes may be found at: http://www.clir.org/pubs/abstract/pub152abst.html

[2] From the Introduction to the Literature Survey by Jerry Persons: http://www.clir.org/pubs/archives/linked-data-survey/part00_01_introduction.html

institutions (e.g. whole universities with their libraries taking the lead) will emit sufficient high quality RDF triples and URIs to complement and extend work underway in numerous museums, publishers, broadcasting agencies, and other agencies in the commonwealth of knowledge. Publication at the source documents of RDF triples and URIs, appropriately reconciled and constantly improved as to the quality of the "facts" and relationships they convey will enable meaningful prototypes of new, efficient, and customizable discovery environments that will speed the processes of generating and promulgating knowledge. Those large and growing stores will also make possible the re-engineering of cataloging and indexing practices that now feed the proliferation of silos of information and meta-information that so limit discovery and thus knowledge generation, teaching, and learning. The full effects of implementing this model in conjunction with Linked Data projects already underway can hardly be predicted other than to suggest that they will be massive and empowering.

Some in the library community fear that the emission of RDF triples and URIs to open stores of Linked Data will further enrich the commercial search engines, catalogs, and indices of the World Wide Web, such as Google. There is similar fear that commercial interests producing indices, abstracts, and fee-based discovery environments will be enriched as well. That is likely so. However, by insisting on open stores of Linked Data, the development of new approaches to discovery for commercial and public purposes, some of them highly specialized, are every bit as likely to be developed. We see this prospect as a true rising tide, lifting all boats, but swamping none.

This plan was devised by Jerry Persons, Philip Schreur, and Michael A. Keller, with significant input from Hugh Glaser. Mimi Calter, and Andrew C. Herkovic provided editorial assistance. Comments, criticism, and suggestions regarding it should be sent to Michael A. Keller (Michael.keller@stanford.edu).

NB: *Structured data* and *Linked Data* are used throughout this paper as synonymous phrases. Furthermore, the phrase "*structured data*" as used herein does *not* equate with *controlled data* in the traditional library sense of that phrase (*i.e.,* controlled vocabularies, use of name authorities, etc.)

Goals

1. Implement an information ecosystem that exploits Linked Data's ability to record and make discoverable an ongoing, richly detailed history of the intellectual activity embodied in all of a research university's academic endeavors and its use of library resources and programs.

2. Design and implement data models, processes, workflows, applications, and delivery services by which academic institutions can create and support pervasive, fully functional capabilities that allow members of the academic community and their compatriots to explore, discover, navigate, access, manipulate[3], and manage the raw materials as well as the finished products of research and scholarship without having to wend their way through the present complex maze of format-driven and vendor-owned silos.

3. Construct an ecosystem based on linked-data principles that draws on the intellectual activity and resources found throughout a research university's programs and its libraries. Use structured, curated representations of these activities and resources to populate a graph[4] of named links. Use this graph to foster the creation of access channels and tools that enhance the quality and reach of discovery, navigation and access capabilities. Pursue designs for these new vehicles and functions that include capabilities whereby use of them by faculty, students, and library staff continuously enhances the quality of the data pool by increasing the density of connections and broadening the scope of relationships throughout the linked-data graph. The objective is a self-improving ecosystem where its effectiveness grows as its use increases.

Scope

The domain of this model comprises the pursuits of a research university's faculty and students. Included in that scope are the knowledge and information resources that a research university creates, acquires, and uses in the course of its scholarship, research, and teaching programs. That range of assets defines the criteria for selecting components of the institution's library collection and service programs for inclusion in the project, creating an academic lens to set the boundary of project activities within any institution implementing

---

[3] Links between controlled data elements will happen immediately. Links between uncontrolled data elements will happen as the data passes through the iterative reconciliation process outlined in Appendix A.

[4] For purposes of this paper, visualize a graph as being a three dimensional array of points. Each point represents a fact about individual fact about a person, place, thing, event, etc. In this array, each fact has one or more links to another fact. Each of these links names the relationship between two facts. For example, the person Samuel Clemens wrote Tom Sawyer. Clemens wrote using Mark Twain as his pen name. Clemens was born in Florida Missouri. That town has lat/log coordinates that locate its position on the planet Earth. Clemens was born 21 April 2010. He married Olivia Langdon. Clara Clemens father was Samuel Clemens … and so on and so forth for all the people and places and things and events that are coming together as a navigable fabric of information and knowledge. For the present discussion, we take a graph to be the array of facts and entities that is navigable via the links that name the relationship between each of the facts in the rapidly expanding web-wide graph of Linked Data.

this model. In addition, this model postulates the generation of RDF triples and URIs as part of the ordinary practices of teaching and research by faculty members and similar figures at other research institutions as well as staff engaged in supporting research and teaching at universities and other research institutions.

Within the compass of this academy/library lens, we will deal with two forms of metadata:

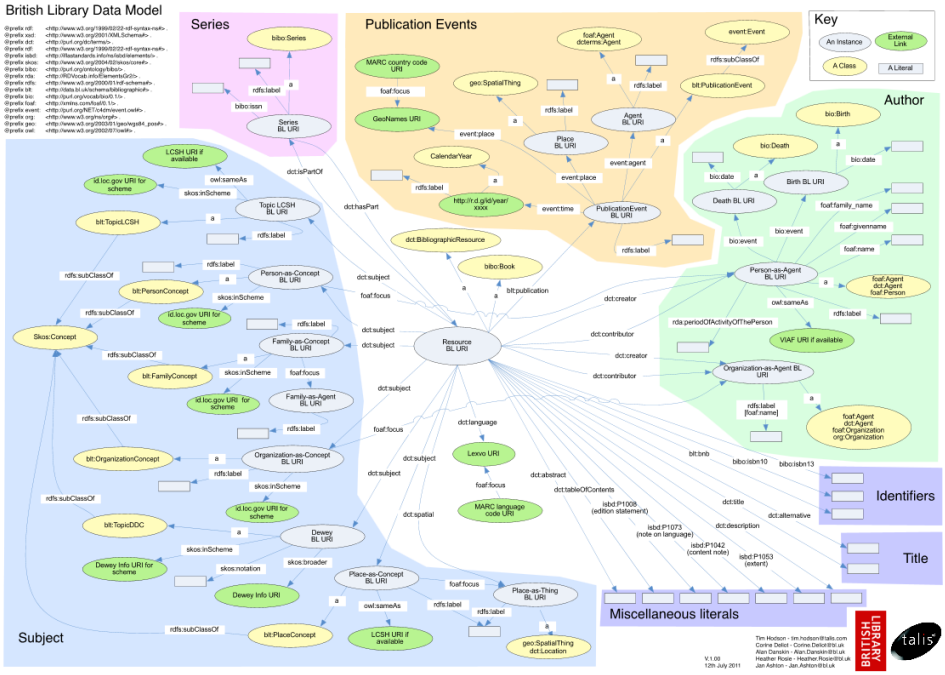
- explicit[5]
  - data that identifies and describes books, articles, media materials, artifacts, research data products, and other forms of content
- implicit
  - course materials, learning objects and syllabi
  - products of citation collection and management tools like Zotero
  - bibliographic links embedded in articles and books pointing to related resources.

[5] Note that some explicit metadata are constructed on widely adopted standards, while others are not.

Approach

This plan adopts today's relatively sparse, loosely woven fabric of fully-open, well-structured, web-friendly forms of structured metadata as the framework on which to build a best first approximation of a generalizable, replicable linked-data ecosystem for supporting the work of students, faculty, and libraries.

Implementation of the plan will require the design, test, and recursively refinement of data models based on the principles of open linked data. It will take as a starting point for this work the recent British Library data model that was developed in consultation with Talis Consulting. (Use the visual link below to examine the full-scale model.)[6]



Doing so will ensure that the resulting model retains the BL's high-level focus and its web-derived, transparent structure for representing facts about people, organizations, places, events, and topics. Such focus represents a marked contrast to efforts based on all-inclusive models that enforce highly structured, deeply detailed and therefore exceedingly brittle representations of physical and digital objects, such as:

- models that closely model traditional cataloging records for books in attempts to replicate the structure and content of such records
- models that delve deeply into various content bearing artifacts' physical/digital characteristics, their history, and the facts and techniques of their creation.

We emphasize that these all-inclusive models such as the two cited here are separate from the model we are describing and out of scope for it.

[6] See http://consulting.talis.com/wp-content/uploads/2011/07/British-Library-Data-Model-v1.01.pdf

Objectives

The proposed model, as well as the attendant processes, workflows, and services that evolve from it, will be considered successful if they:

1. are fully open, *i.e.,* all data and services are licensed as CC0 or equivalent
2. rely on general-purpose web-based protocols, schemas, tools, and processes
3. remove the strictures of format-driven silos
4. decompose records into a fabric of paths across navigable statements of fact
5. break down IP constraints by focusing on statements of facts, rather than records
6. lend themselves to being improved in breadth, quality, and density as use increases
7. help spread academically validated links and content throughout the web of data
8. act as self-improving ecosystems driven by community activities

Our broader objectives in producing the model are:

1. to allow an academic institution, its faculty and students, and its libraries to operate both as full-fledged participants, and as active change agents in shaping the emergent web of data
2. to bridge today's multiple, fractured, un-linked, and uncoordinated streams of services and resources in order to move libraries into the well-structured, web-transparent, linked-data environments that are emerging.

Specific objectives for this implementation are:

3. to implement the model in a way that adds no incremental cost to library technical and public services
4. to implement the model in a way that can be reproduced with little coordination among institutions engaging the model.

Environment

This plan takes advantage of the confluence of destabilizing factors at work in today's research university and library environments. These factors include:

- turmoil in many components of the scholarly communications food chain
- rapid if not exponential growth in interdisciplinary scholarship and research
- continuing pressure on library programs to increase efficiency and reduce costs
- rapid evolution of basic components of library metadata environments (RDA, MARC)
- demand for access to non-traditional resources (e.g., finding aids, images) within traditional catalogs
- internationalization of metadata standards and authorities
- drive to freely accessible and open data
- proliferation of competing discovery services
- increase in "semantic" services provided within individual publishers' silos
- recognition of the value of information and learning objects in a variety of formats that formerly were not visible, not shared, and/or valued only by their creator and his/her immediate audience (students, post-docs, immediate colleagues/collaborators)
- proliferation of institutional and disciplinary repositories, themselves hindering effective discovery of usable information objects and ideas

The confluence of change agent in the present environment makes possible the reshaping of the methods use by research universities in managing intellectual resources, while at the same time making major improvements in library programs and services that deliver those resources back.

Products

Implementation of this model in one or more institutions will produce a replicable exemplar for the changes that can be made in the creation and use of information/knowledge resources and services though application of tools, methods, processes, and workflows based on open, well-curated structured data.

Implementation of this model allows an academic institution, its faculty and students, and its libraries to operate both as full-fledged participants, as well as active agents for change in shaping those aspects of the emergent web of data that will impinge on the programs of research universities and their libraries.

Implementation bridges between today's multiple, fractured, un-linked, and uncoordinated streams of services and resources and the well-structured, web-transparent, linked-data environments that are emerging on the near horizon.

We believe implementation of the model will results in a zero-sum increment to today's library budgets for technical and public services.

Implementation should be replicable with no more than modest coordination among institutions engaging this model.

Architectural Concepts

Target

Pursuit of one or more user interfaces, "killer apps", for linked data was an oft recurring topic of discussion throughout the Linked Data Workshop at Stanford in late June, 2011. It is also a recurring thread in every venue associated with linked data. Indeed, the question remains pertinent: why pursue a linked data approach in the present case?

There are examples of tools that provide glimpses of what will be possible as the fabric of well-structured data continues growing toward the early stages of its of web-wide maturity. Analysis of one example, LinkSailor, is provided below.

But the case for using linked data principles is less a matter of new or even revolutionary types of interfaces (*i.e.*, killer apps), than it is a matter stepping across the tectonic fault line that separates today's metadata processes and workflows from the linked-data driven capabilities by which institutions and organizations (in this case research universities and their libraries) could go about managing their knowledge and information resources.

The rationale for adopting linked-data principles for this project:

Present-day metadata workflows and processes are rooted in descriptions and topical analysis of artifacts that transmit content in a variety of physical and digital formats (books, articles, media, databases, learning objects, etc.).

Linked data does offer useful enhancements to the descriptive aspects of metadata for resource description and management:

- *records decomposed into statements of fact with strong identifiers*
- *reconciliation of connections among such facts that cross format and genre boundaries*
- *links that tie facts together into a web-wide graph of connections.*

While of measurable value, these improvements do not support making wholesale changes in present-day practices and systems.

What does warrant our attention, and does support making the institution-wide changes proposed for this project, is linked data's ability to record and make discoverable an ongoing, richly detailed history of the intellectual activity embodied in all of a research university's academic endeavors and in all the academy's use resources and programs of its research libraries.

Linked Data methodology has the capability to track and make use of how knowledge and information resources were and are being used in research, scholarship, and teaching, as well as in library service and collection programs.

In addition, the Linked Data model has the capability to navigate through and across the boundaries of the active, every growing fabric of academic disciplines via links that include:

- *citation maps that weave together publications supporting data sets*
- *pointers reaching inside content vehicles (links based on book indexes)*
- *course materials (syllabi, reading lists, examinations, learning objects, videos of lectures, slide decks, data)*
- *products of day in, day out research activities (Zotero, RefWorks)*
- *activities/products of the library's reference, collection, and services programs*
- *links capturing the findings and musings of gifted/prominent faculty/researchers/teachers on their special interests*
- *paths that bridge among topical ontologies, taxonomies, vocabularies, etc.*
- *links that wend their way across institutional boundaries*
- *connections that tie related facets of disparate disciplines together*

A result of adopting the Linked Data approach is the capacity to lend structure to every form of publication across the entire range of university and library activities:

- *in/formal documents, presentations, seminars, conferences, exhibits, …*
- *information embedded in web pages across an entire [university].edu domain*
- *all manner of materials created in the course of library service, exhibit, collection, and other programs.*

The resulting capability will capture an ongoing, structured record and build from it a navigable tapestry that weaves together every facet of how knowledge and information resources are used an institution's research, teaching, and scholarship.

> *And that adoption will involve every aspect of a research library's programs, and every member of its staff, in creating, curating, and publishing the ongoing, academy-wide intellectual history of their institution to a web-wide audience.*

If we take as a given the existence of a project to create and manage linked data at a scale comparable to the implications of the forgoing rationale, one might expect to see a complex array of tools and applications playing important roles. Putting a high-level scan of such components off until the following section on **Infrastructure**, here is the promised look at one precursor of the type of tools that will be used to explore the fabric of links in a university-wide pool of structured data with URIs providing strong identifiers for every RDF triple.

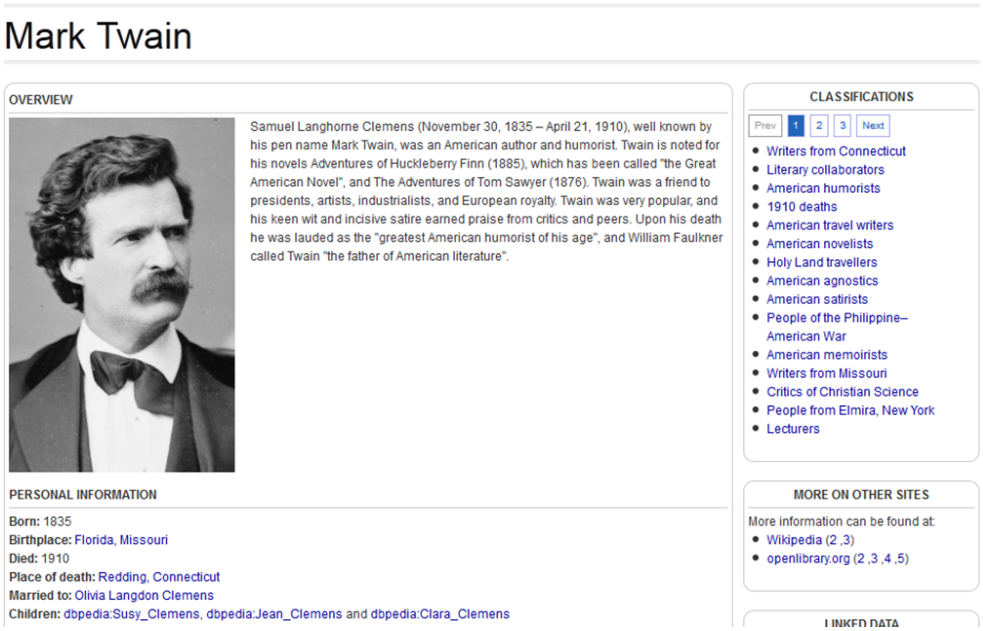
Interactive elements

LinkSailor is a service built by Talis using the capabilities of the Talis Platform, their structured-data development and processing environment (indeed, the Platform provides the infrastructure behind their emerging data-market product known as Kasabi). The LinkSailor's structured-data engine traverses the fabric of links from a selected set of environments to assemble a well-ordered display of facts and pieces of information about a selected name or topic. This is a dynamic process—one that sees when content or links get added, changed, or deleted, modifies the resulting display to follow those changes. In fact, depending on the speed of one's internet link the page image may move around in somewhat jerky fits and starts as URIs are resolved and their textual labels replace the lengthy URL strings.

To get a sense of this sevice, begin with the LinkSailor home page[7] as the starting point, and select Mark Twain from the menu:

[7] See http://linksailor.com/nav

The interface looks a lot like most web sites of this type (Wikipedia-ish). There is a picture of Twain/Clemens, a bit of biographical text, some facts about his life (birth and place, death and place, wife and children). To the right (*Classifications*) is a list of topics that Wikipedia (via its data export known as DBpedia) associates with Twain. These links take you to appropriate articles in Wikipedia.



In the box below this (*More on other sites*) there are links leading to more information about Twain at locations that include Wikipedia, openlibrary.org, and the New York Times. At the Times site, there is a column to the right labeled: *Samuel Clemens Navigator: A list of resources from around the Web about Samuel Clemens (Mark Twain) as selected by researchers and editors of The New York Times.* In the future ecosystem predicted by the elevator pitch, this type of structured-data navigation environment would have traversed the links posted by the NYT, and included direct access to entries (among others) for:

- *Mark Twain Papers and Project*, Bancroft Library, UC Berkeley
- *Mark Twain and His Times*, University of Virginia
- *Mark Twain interactive scrapbook*, from PBS
- *Lionel Trilling on Mark Twain* (NYT, 1946)

Furthermore, an environment driven by linked-data representations of the history of local academic activities and of work created by library programs would have provided an alert to the English Department's Fall Quarter course: Mark Twain and American Culture (plus a number of other courses with syllabi and reading lists from the past several years). Also, the Library's topical guide for resources associated with African colonial history that points to Twain's *King Leopold's Soliloquy* would have been revealed. This work is referenced as part of a web site: Mark Twain's Anti-imperialist Writings: a Guide to Online Resources, And to flesh out a reliable path to an online copy of the *Soliloquy*, there would be an alert to the reserve reading list for English 320, Practical Criticism via a humbolt.edu link (Humbolt State University, CA)—a link pointing out that:

the copy of this text that used to reside at Jim Zwick's Mark Twain's Anti-Imperialist Writings: A Guide to Online Resources is gone, but a PDF facsimile of the original is available at the American Museum of Natural History's "Congo Expedition: 1909-1915" site.

This short demonstration illustrates the potential scope of resources that would be made accessible via a cohesive, structured-data ecosystem once the breadth and depth connections can be expanded to include the whole of an academic institution's intellectual endeavors, and once such coverage begins to include large numbers of similar pools of information generated by sister institutions in the US and around the globe. The resulting capabilities, seen in early sketch form in LinkSailor, offer the prospect of leaving behind the arduous tasks of hunting and pecking through the quirks of multiple local and vendor interfaces in combination with weeding through massive web search-engine responses for the few tidbits that relate to work on a given research topic.

For those with a taste for exploring what is under the hood in the structured-data engine behind LinkSailor, return to its home page, and click on the **| Show data |** tab in the upper left corner of the page. Scroll down to the grey-highlighted URL above rdfs:label Mark Twain (http://linksailor.com/nav?uri=http%3A//semanticlibrary.org/people/mark-twain).
Click this URL, and theTwain page seen earlier returns. Click on the **| Show data |** tab in the upper left corner of this Twain page. You will see LinkSailor go through the process of resolving the 250+ lookups that bring back information spread out across the threads of structured data associated with Mark Twain in the Talis Platform's structured data pool. A quick scan will reveal references of varied types from across the linked-data cloud:

cc:attributionName – CC BY provenance statement regarding data from Freebase
nyt:topicPage – the aforementioned reference to the NYT landing page for Twain
dbo:[various values and text] – from the DBpedia export of Wikipedia content
dbo:birthplace and rdfs:label and foaf:name – for Florida, Missouri
geo:lat and geo:long for the town's position on the globe
db:genre – the aforementioned topical headings from DBPedia/Wikipedia
dbp:wordnet_type – a URI linking the Twain name to wordnet's writer/noun statement
dct:alternavtive – Dublin Core reference to the name form Samuel Langhorne Clemens
dct:subject – topics, also expressed in the SKOS schema
bio:[elements] – expressions of biographical facts using a biography schema
fb:[elements] – Freebase statements about Twain's works, books, etc.

-- in association with information regarding King Leopold's Soliloquy
fb:type.value.value -- Freebase keys for the work
from widipedia: King_Leopold$0027s_Soliloquy
from Freebase: king_leopolds_soliloquy
fb:type.object.name
from Freebase: King Leopold's Soliloquy
rdf:type -- http://rdf.freebase.com/ns/book.book
-- http://rdf.freebase.com/ns/book.written_work
xhtml:license
-- http://creativecommons.org/licenses/by/3.0/
owl:sameAs

-- http://dbpedia.org/resource/King_Leopold's_Soliloquy
fb:book.written_work.previous_in_series
-- http://rdf.freebase.com/ns/en.a_dogs_tale

From this highly selective extract of the data behind LinkSailor's take on Twain, it is easy to see how much information is embedded in a structured data ecosystem, even at this very early stage of maturity for web-wide linked-data environments. When the density of the graph's fabric and scope of growing coverage from academic institutions comes into play, the capabilities of discovery, navigation, and access tools that are the children and grandchildren of LinkSailor and its siblings will need to provide all manner of personalization alternatives, *e.g.*, capabilities allowing one to filter and select for relevant resources and information from a wealth of alternatives. As one colleague noted at the Stanford Linked Data Workshop, the problems of scale will not, in fact, be problems … they will be demonstrable measures of success.

For a slightly orthogonal take on new interface ideas, see the Code4lib email archive for an array of messages that provide a quick, up-to-date scan of visually based search interfaces that are coming into play. One eye-catching mockup comes from Harvard as a contribution to this fall's DPLA (Digital Public Library of America) proposals. Dubbed LibraryCloud, it is an alpha implementation of metadata services that aggregate extracts from traditional library metdata records with a variety of facts related to circulation, reader reviews and ratings, social interactions, and other types of information. The interface that takes advantage of this array of data is called ShelfLife, an interesting collection of ideas and approaches that have merit as a sampling of capabilities that could be built over aggregated pools of linked data (suggestion: the tour provides a bit less attitude and quite a bit more information).

Another approach can be seen in the Beta of Microsoft Academic Search's Visual Explorer. To activate the maps, search for an author in the upper-left-hand box. Alan Jones as a search argument brings up a gent from Indiana University. From there one can look at a co-author graph, co-author paths, and a citation graph. All early days (a bit too heavy on graphical design, and a bit light on content, perhaps), but an indication of what interfaces based on structured data might accomplish. This example also illustrates the need for strong identifiers, URIs associated with RDF triples, for numerous false relationships appear in it.

Ecosystem

As noted earlier in the introduction, an essential initial step in this work will be designing and iteratively refining a data model that is flexible enough to support processing, management, distribution, and access services for the included span of academic and library resources. This model must provide a carefully balanced combination of:

- detail that suffices for back-of-the-house processing and management work
- flexibility that can accommodate the rapidly evolving conventions of structured data
- transparency that allows it to converse seamlessly with web-wide tools and services
- complexity that grows to support more densely woven tapestries of navigation and discovery paths as the quality and depth of structured data improves

We believe that the British Library Data Model represents the best first approximation of the requisite framework for our project. We will work with our colleagues at the BL and with the linked-data specialists at Talis Consulting to understand fully the strategic and tactical thinking that lies behind the BL model. Working in concert with those agencies, we will consult with others who have a track record of success working with linked data (BBC, data.gov.uk, Hugh Glaser, etc.—for more details about these and related endeavors, see the Richard Wallis presentation at the Talis *Linked Data and Libraries, 2011*).

With a well-vetted data model in hand, we will then pursue design of an environment that will support the objectives of our project. Rather than starting from the inside and working our way out (*i.e.,* starting with back-room processes/data and working our way out toward discovery/delivery services and interfaces), we propose starting in the middle and pursuing improvements in both directions (see the component outline below for further details). Many factors make such an approach necessary:

1. The long-lived infrastructure that has shaped library metadata processing and services since the 1970's will undergo a substantive revision as the Library of Congress and its national and international partners work their way through its proposed transition to A Bibliographic Framework for the Digital Age.

2. Linked data itself is in a state of considerable flux as it moves out of its development in academic environs to become a productive subset of still-distant semantic web technologies. A telling example of the range of competing opinions about what constitutes "good" structured data can be had in a review of schema.org's appearance on the web-of-data scene in June, 2011 (summarized in the Survey developed for the Stanford Linked Data Workshop). Suffice it to say here for our purposes that we will see a continuous flow of changes in what's needed to be an active participant in the development and promulgation of linked data.

3. Academic publishing in general and the metadata that underpins discovery and access services for the resources that fuel academic programs are under pressure to change based on a variety of structural, cultural, and economic forces. Scholarly and professional societies who publish as well as for profit academic publishers wish to aggregate content they process and distribute. They also want to aggregate and make discoverable information about conferences, career building & employment opportunities, collaboration, commercial and other services supporting research, and funding programs. Some of them are developing Linked Data programs, albeit not ones that emit open and freely usable RDF triples and URIs that provide actionable and constantly updated links in support of scholarship, professional practices, continuing education in the professions, teaching, and learning. All publishers are seeking compelling services that tie users to them on an ongoing basis.

4. On the metadata front, open data policies have begun to take root with remarkable results. More than 40 national libraries in Europe recently voted to support an open data policy for their bibliographic records.

5. Many research libraries have already broken the link between their book cataloging systems and the engines used to deliver discovery, navigation, and access services.

Out of necessity, they've had to build various metadata creation and management environments to accommodate materials that cannot be managed as a part of traditional book-cataloging workflow and processes

As work proceeds from modeling toward planning for workflows, processes, and services, we expect to focus on capabilities that implement an environment that can adjust to ongoing levels of active change--both on the in-house side of things, as well as in the data structures and requirements for discovery, navigation, and access services, and also in the communication channels that allow the local environment to interact with web-wide environs. This means that our proposed mid-level ecosystem for structured data will need to accommodate and interact successfully with:

1. the traditional library processing environments as they morph toward support for revised cataloging standards (RDA) and new vehicles for sharing the work of building metadata for research collections (Library of Congress and related projects)

2. an array of varied metadata and content management engines, some from primary and secondary scholarly publishers, created to deal with the resources that fall outside the capabilities and policies of present-day cataloging systems and environments

3. an evolving set of tools and infrastructure that will emerge as the mining of various forms of implicit metadata begins to have an impact on the capture, management, and sharing of structured data

4. increasing amounts of more finely grained structured data about members of the academic community, their activities and research, and the (un)published products of their work

5. ongoing refinements in the level, quality, breadth, and complexity of structured data that flows in from web-wide services and resources

6. the evolution of discovery, navigation, and access vehicles from the early stages of layering extant interfaces over emerging flows of structured data, and eventually through entirely new types of discovery and navigation tools.

Sustainability

Of central importance to the planning process will be discovering how to make this project move steadily toward a self-sustaining state. Evolution from today's state of affairs through the matrix of changes that are forthcoming must proceed along paths that convert resources that now support in-house metadata creation and management into capabilities that can sustain management and delivery of well-structured data for much broader range of materials, programs, people, and services.

Achieving this goal will require active, widely based consultation throughout the academic community. Broadly cast outreach efforts must begin at the inception of the project, and go hand in hand with conceptual design and strategic planning. One of the key elements of

success for this project is creating an ecosystem of structured data that lends itself to curation and improvement by the members of the communities that it serves. Ensuring the scope and levels of activity for that contribution, *i.e.*, building crowdsourcing in as an essential component of managing the new structured data ecosystem, must be a fundamental imperative throughout the project.

The second and equally important aspect of sustainability is funding. We view this project as a bridge between today's tools, practices, workflows, and systems that connects present capabilities with those that will support the structured-data environment outlined for this project. Once the transition is accomplished, resources and staff and budgets that support today's environment must suffice to support the new environment. The objective in terms of funding is a zero-sum budget increment at the project's conclusion.

Project phases

In concert with the twin focal points of academic pursuits and library programs, this effort will consist of two phases in which all activities derive from the aforementioned data model.

Phase one: Work within the university/research community will include efforts to identify and populate appropriate structured data representations for

- the people and organizations that make up the academic community
- the publications, reports, proceedings, and other content created by them.

We note that a growing number of web-site creation and management tools provide access to capabilities that bring provision of structured data closer to becoming an everyday part of contributing content to the web. Drupal's RDFa support in its core modules is one example. Another is the well documented and steadily evolving implementation of microdata provided by schema.org.

From the library perspective, work will include identifying and gathering

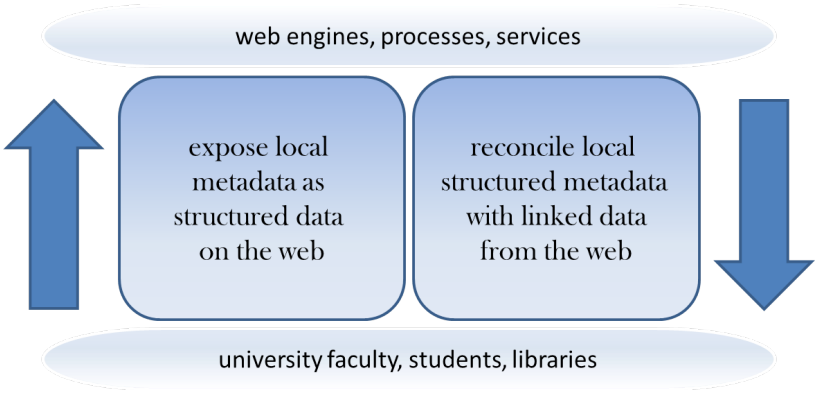
- the various sources and pools of metadata associated with library resources
- the sources and pools of implicit metadata embedded in courses, in Zotero-like tools, and in the bibliographic apparatus found in books, journals, dissertations, etc.

Once the raw materials are identified and assembled, the next steps will include transforming the resulting pool of metadata into an institution-wide set of linked-data statements. That collection and processing environment (a prototype for an eventual institution-wide ecosystem) will need to support creation and management of RDF statements generated from

- extant pools of metadata
- newly created metadata for traditional materials and forms of publication
- updates generated by and fed back into extant systems
- new processes and workflows designed to capture, analyze, and manage structured metadata extracted from content as it is newly minted by members of the academic community
- ongoing analysis of streams of implicit metadata.

The resulting pool of structured-data statements will feed into an array of web-wide engines, processes and services including projects and services supported by Freebase,

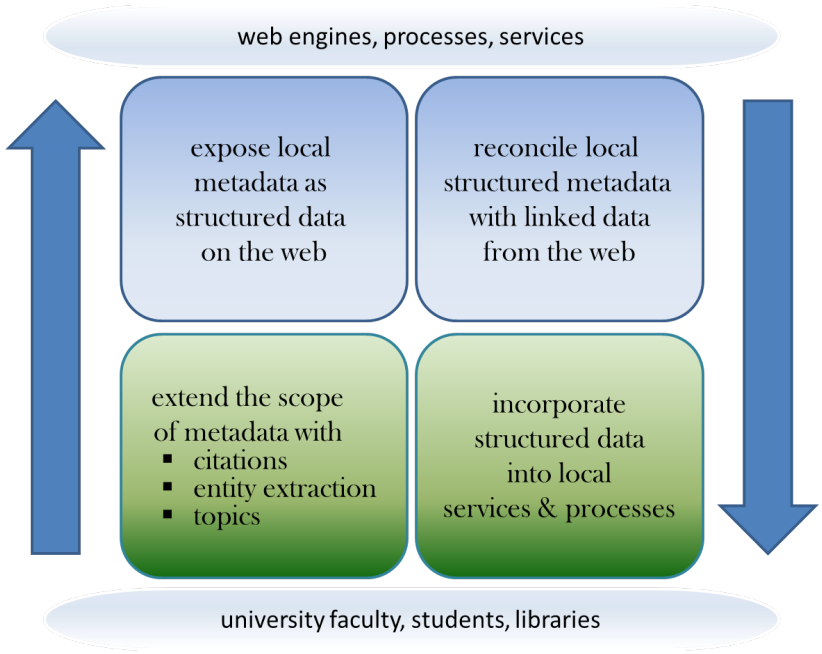
Seme4 and <sameAs>, and the Talis Platform (see the section on **Infrastructure** for more details). As seen in the illustration below, the objective here is for those widely-scoped, web-based environments to feed extensions back into the pool of locally generated statements--extensions that include reconciliation of people, organizations, places, events, topics, and other entities found in the expanding fabric of well-curated linked data as it spreads across the emergent web of data. We anticipate both local and global stores of RDF triples and related URI identifiers with associated services.



Phase two: Work will focus on expanding the depth and density of the local pool of structured data that is fed out into web-wide services. Work will explore the

- extraction of citation data from books, articles, and other artifacts
- value and effectiveness of entity extraction from textual content
- processes and prototypes for transforming and interlinking extant topical schemas, vocabularies and taxonomies

In concert with additions to the outbound flow of information to web-based reconciliation engines, efforts will also address the all important aspect of creating an extensive prototype for the delivery phase for new institution-wide knowledge and information processes and services. This environment will be driven by an increasingly rich, completely open, highly interlinked, fully replicable ecosystem. One that is continuously curated, expanded, and improved via the very processes and services by which the academic community makes use of these new types of resources, tools, and capabilities.



Beyond the conclusion of this project, say something on the order of five or so years later, the ecosystem generated by the explicit work of building and running a continuously expanding linked-data prototype from multiple different university or research institutions as described in this model will have evolved toward becoming the new norm for managing and using the intellectual resources that fuel research, scholarship, teaching, and learning. With continued growth in the density and breadth of the linked-data graph (especially when links begin to permeate the boundaries between multiple disciplines and bridge the gaps between institutions) will come rapid increases in the number of tools and types of capabilities for contributing to and drawing on a web-wide pool of knowledge and information. When the internal processes and systems start to use linked data as in the lower left box of the diagram, as we expect them to over the years, the old IT systems will no longer need to be maintained, and so the publication of linked data will no longer be an additional cost.

Just as no one today thinks twice about creating sophisticated documents that combine textual, graphic, and video content (aside from the gearheads and geeks who make the technology work), no one in the near-term, five to ten year future will have to worry about "making Linked Data" or "building RDF triples" or "writing triple-store database queries". Precursors of the simple-to-use and increasingly robust capabilities that will mask the details of what makes Linked Data work already exist in many venues. LinkSailor is a promising approach to discovery and navigation. Drupal, an open-source environment for managing sophisticated, data-rich web sites, included linked-data capabilities in the core modules of its version 7 release. The community is actively at work adapting that new environment to the ongoing evolution of the linked-data world, as a scan of Dupalcon's program Spring 2012 in Denver illustrates. Talis has built wide adoption of its Aspire product--marketing based on a service that helps teachers, students, and their parent institutions manage broad access to learning resources. A quick scan of the product's homepage includes capabilities like *add resources from leading providers with just a couple of clicks … rich metadata, library linking and acquisitions alerting all taken care of – no form filling*

required. On the surface, Aspire is a set of tools and capabilities that meet specific day-to-day needs of teachers and students. Under the hood, there's a full-featured linked-data environment, one that allows Talis to offer capabilities through Aspire that include *search and discover a world of learning resources, organized by discipline and topic and focused on UK HE … browse recommendations based on actual usage by peers in taught courses across UK Universities.*

As tools and capabilities like these mature and spawn their successors, they will become the human interfaces that mask all the complex plumbing that is needed to support building and managing and using web-wide pools of structured data. They will allow and encourage increasingly high levels of participation by all members of the academic community in adding to and refining the web-of-data as a normal part of the academy's day-to-day use of knowledge and information resources. Indeed, they will help foster an ecosystem that is continuously curated, expanded, and improved via the very processes and services by which the academic community goes about the work of building and using intellectual resources.

Project components

Infrastructure

Having set the bar for this project at the level of delivering capabilities that can record and make discoverable the full history of a research university's and its libraries' intellectual program activities, what are the components of an infrastructure that can accomplish our aims, and what are the models and projects that we can turn to for guidance and tools? One such model is the British Museum's ResearchSpace.

[From the project's home page,] *ResearchSpace is an Andrew W. Mellon Foundation funded project aimed at supporting collaborative internet research, information sharing and web applications for the cultural heritage scholarly community. The ResearchSpace environment intends to provide following integrated elements;*

- *Data and digital analysis tools.*
- *Collaboration tools*
- *Semantic RDF data sources*
- *Data and digital management tools.*
- *Internet design and authoring tools*
- *Web Publication*

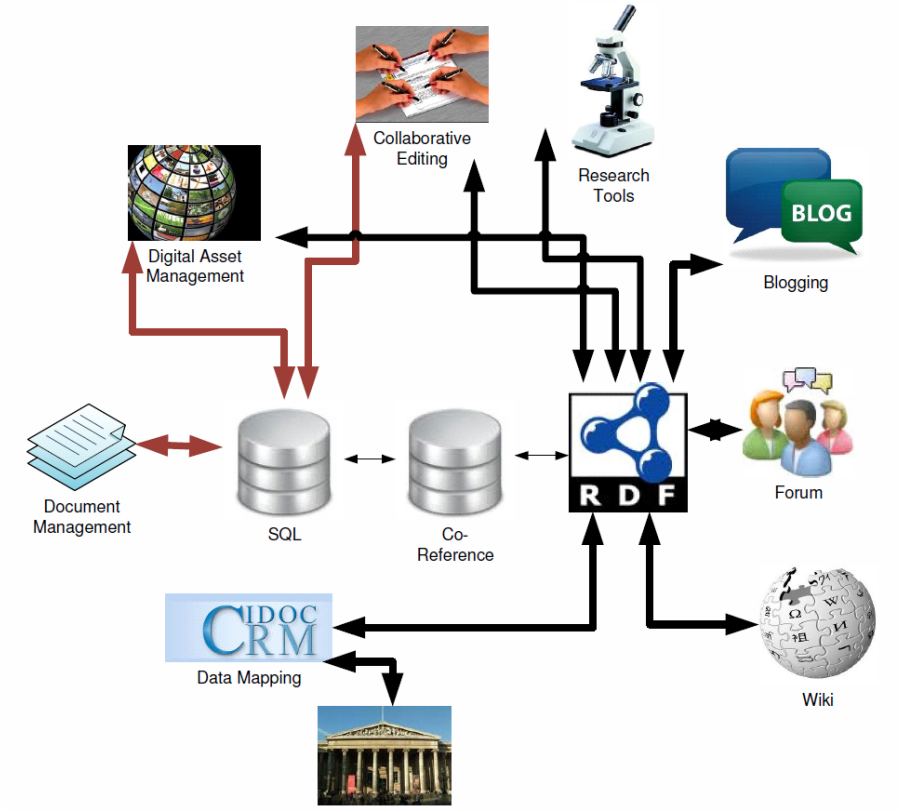
With products of their investigative and planning work starting to appear over a year ago, this project has done a great deal of the intellectual and IT-related spade work required to address the issues facing the project under discussion here. Allowing for the expected differences between their focus on museum practices, content, and research, plus attendant data models and conventions, the commonalities between what they are pursing and we are proposing are numerous and well-suited to our goals and objectives. Those commonalities are the structured data arising from the full rang of resources generated and used by scholars and others in academic pursuits, the intersection and overlaps of our data models, our commitment to emitting Linked Data in open stores for open and free use, and by our devotion to dramatic improvements in discovery environments. For example, Dominic Oldman's presentation this fall at the Yale Center for British Art (The Future of Research) provides a nuanced and compelling case for the adoption of a structured-data ecosystem to support the British Museum's (and those of sister institutions in the project) varied needs—needs that are very much akin to recording and playing back the full history of a research university's and its libraries' intellectual program activities.

Via their ResearchSpace Business Requirements & Specifications (v.2, May 2011) we have access to the analysis and planning that lies behind specifications for key components of a structured-data environment:

- collaborative content management
- social networking tools
- document/asset management
- research and collaborative editing tools
- data stores and data synchronization mechanisms

Allowing for the differences between museums and research university/library programs, we can learn much from The ResearchSpace's groundbreaking work. While our project's

infrastructure will not be a carbon copy of theirs, the loosely coupled and coordinated array of elements in their component model resonates with what we will need to create



Schematic Snapshots

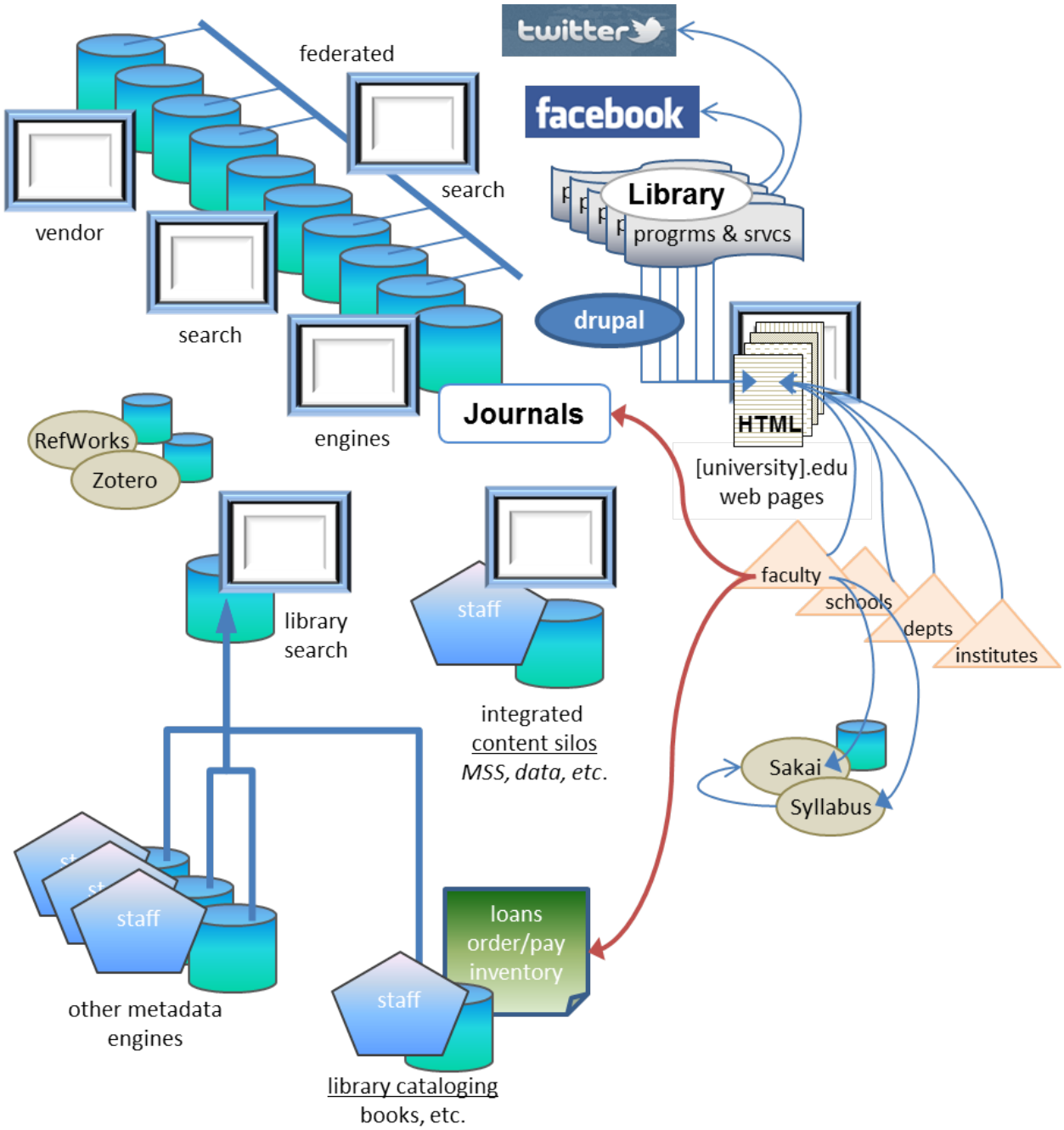
What follows is a set of schematic sketches accompanied by brief commentary aimed at outlining phases and components of the project.

1. Current situation [full-page image]

Moving clockwise around the sketch, we see

- library programs and services (including collective ones based on OAI and similar repositories -- ePrints, Dspace, etc)
  - some have connections to various social tools and environments
  - most have a presence in the [university].edu web space
  - delivery of some programs/services makes use of a CMS (Drupal, etc.)
- faculty, schools, departments, institutes, etc.
  - faculty publish books, journals, reports, conference contributions, etc.
  - they post some portion of that material on web sites (personal, school, etc.)
  - schools, departments, institutes, etc. have a substantive web presence
  - faculty make use of course mgmt. tools & services (Syllabus & Sakai at Stanford)
- library acquisitions, cataloging, circulation, and inventory management
  - traditional library management system processes and services
- other metadata engines
  - these environments build metadata for materials not suited to catalog control,
    much of such content is digital, much of it requires extended metadata (preservation, provenance, formats, etc.) that doesn't fit cataloging schemas, policies, and systems
- library search
  - many research libraries have broken the link between the OPAC packaged with their LMS and moved on to engines like Blacklight to provide search access that spans metadata from cataloging and other metadata sources
- tools used to support the day-in-day-out work of scholarship and research
  - RefWorks (citation capture tool) and Zotero are cited as typical of the type
- vertically integrated content silos
  - at Stanford these include medieval manuscripts, social science data, EADs, etc.
- journal literature
  - one commonly finds some combination of vendor search environments and federated search, plus some form open URL resolver (SFX from ExLibris in the Stanford case)

Current Situation



2. Phase-one components [full page image]

Moving clockwise around the sketch, we see

- library programs and services
  - the RDFa component refers to the Drupal 7 linked-data module
  - intent is to begin capturing content, e.g., subject and collection guides
  - this to demonstrate linked-data capabilities to campus Drupal community
- faculty, schools, departments, institutes, etc.
  - mine university ID-card data for faculty names, departments, etc. (some 2,000 faculty and ca. 1,000 schools, depts.., etc. at Stanford), separating public from private data on individuals
  - mine various metadata pools for faculty's articles, books, reports, etc. (see Materials & URIs below for details about this effort)
  - include connections to/from ORCID, and possibly VIVO if appropriate
- linked-data *transformation workflow*
  - this component is a place holder for the workflows, methods, and processes that will produce the local pool of linked data, see first-pass level of planning by one of the workgroups at the Stanford Workshop
  - we expect to expend considerable effort on this component, working in consultation with a number of our partners (the Metadata staff at the British Library, Talis Consulting, Hugh Glaser and his colleagues at Seme4), plus members of the British Museum team and colleagues who are working to shape the structured data work that supports Europeana.
  - as noted above, this planning must provide for an evolving ecosystem, one in which:
    - the linked-data model is undergoing continuous refinement
    - the scope, processes, and vehicles for crowd-sourced input by the academic community and all contributors to library programs and services move from infancy through various stages of maturation

Phase one components