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:

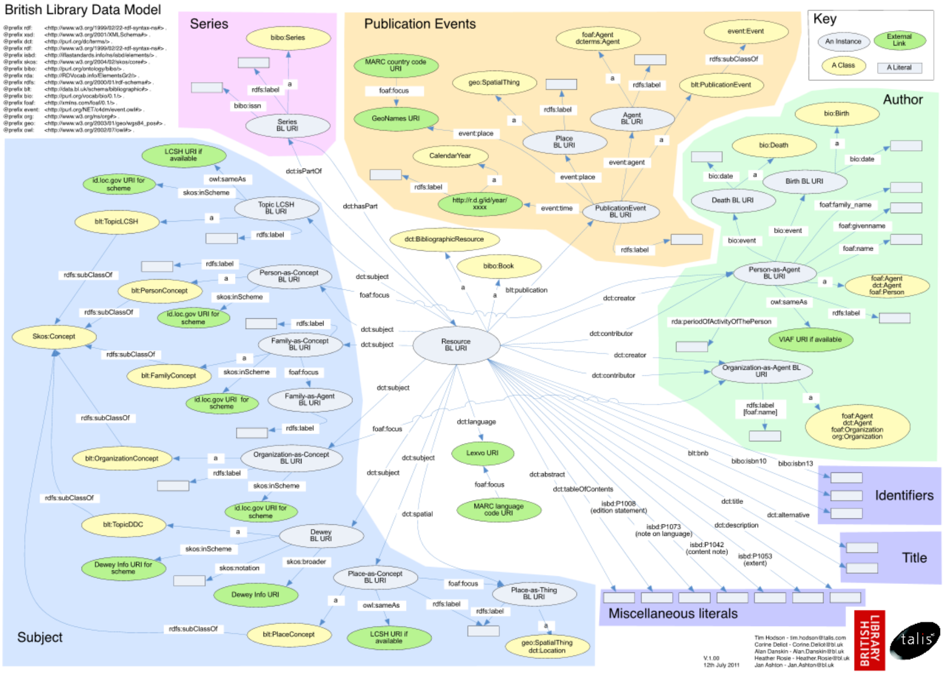
- explicit[5]
  - data that identifies and describes books, articles, media materials, artifacts, research data products, and other forms of content
- implicit
  - course materials, learning objects and syllabi
  - products of citation collection and management tools like Zotero
  - bibliographic links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


British Library Data Model
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix isbd: <http://iflastandards.info/ns/isbd/elements/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix rda: <http://RDVocab.info/ElementsGr2/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix blt: <http://data.bl.uk/schema/bibliographic#> .
@prefix bio: <http://purl.org/vocab/bio/0.1/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix event: <http://purl.org/NET/c4dm/event.owl#> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
Series
bibo:Series
rdfs:label
a
bibo:issn
Series BL URI
Publication Events
MARC country code URI
foaf:focus
GeoNames URI
event:place
geo:SpatialThing
CalendarYear
rdfs:label
http://r.d.g/id/year/xxxx
event:time
Place BL URI
foaf:Agent dcterms:Agent
Agent BL URI
event:agent
event:Event
rdfs:subClassOf
blt:PublicationEvent
PublicationEvent BL URI
Key
An Instance
External Link
A Class
A Literal
Author
bio:Birth
bio:Death
bio:date
Birth BL URI
Death BL URI
bio:event
foaf:family_name
foaf:givenname
foaf:name
Person-as-Agent BL URI
foaf:Agent dct:Agent foaf:Person
owl:sameAs
rdfs:label
rda:periodOfActivityOfThePerson
VIAF URI if available
Organization-as-Agent BL URI
rdfs:label [foaf:name]
foaf:Agent dct:Agent foaf:Organization org:Organization
dct:isPartOf
dct:hasPart
dct:BibliographicResource
bibo:Book
blt:publication
Resource BL URI
dct:creator
dct:contributor
dct:subject
foaf:focus
dct:language
Lexvo URI
MARC language code URI
dct:spatial
dct:abstract
dct:tableOfContents
isbd:P1008 (edition statement)
isbd:P1073 (note on language)
isbd:P1042 (content note)
isbd:P1053 (extent)
dct:description
dct:title
dct:alternative
blt:bnb
bibo:isbn10
bibo:isbn13
Identifiers
Title
Miscellaneous literals
LCSH URI if available
id.loc.gov URI for scheme
skos:inScheme
Topic LCSH BL URI
blt:TopicLCSH
rdfs:subClassOf
Person-as-Concept BL URI
blt:PersonConcept
Skos:Concept
Family-as-Concept BL URI
blt:FamilyConcept
Family-as-Agent BL URI
blt:OrganizationConcept
Organization-as-Concept BL URI
blt:TopicDDC
Dewey BL URI
Dewey Info URI for scheme
skos:notation
skos:broader
Dewey Info URI
Place-as-Concept BL URI
Place-as-Thing BL URI
blt:PlaceConcept
geo:SpatialThing dct:Location
Subject
Tim Hodson - tim.hodson@talis.com
Corine Deliot - Corine.Deliot@bl.uk
Alan Danskin - Alan.Danskin@bl.uk
Heather Rosie - Heather.Rosie@bl.uk
Jan Ashton - Jan.Ashton@bl.uk
V.1.00
12th July 2011
BRITISH LIBRARY
talis

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.

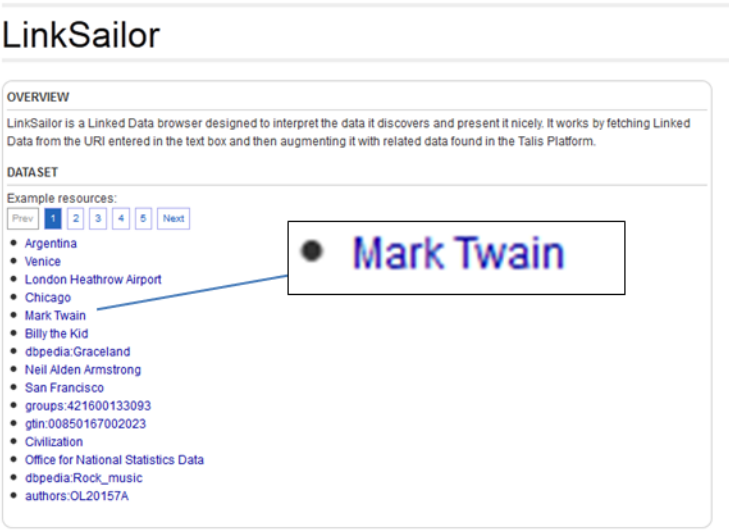
> *And that adoption will involve every aspect of a research library's programs, and every member of its staff, in creating, curating, and publishing the ongoing, academy-wide intellectual history of their institution to a web-wide audience.*

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:



LinkSailor

OVERVIEW

LinkSailor is a Linked Data browser designed to interpret the data it discovers and present it nicely. It works by fetching Linked Data from the URI entered in the text box and then augmenting it with related data found in the Talis Platform.

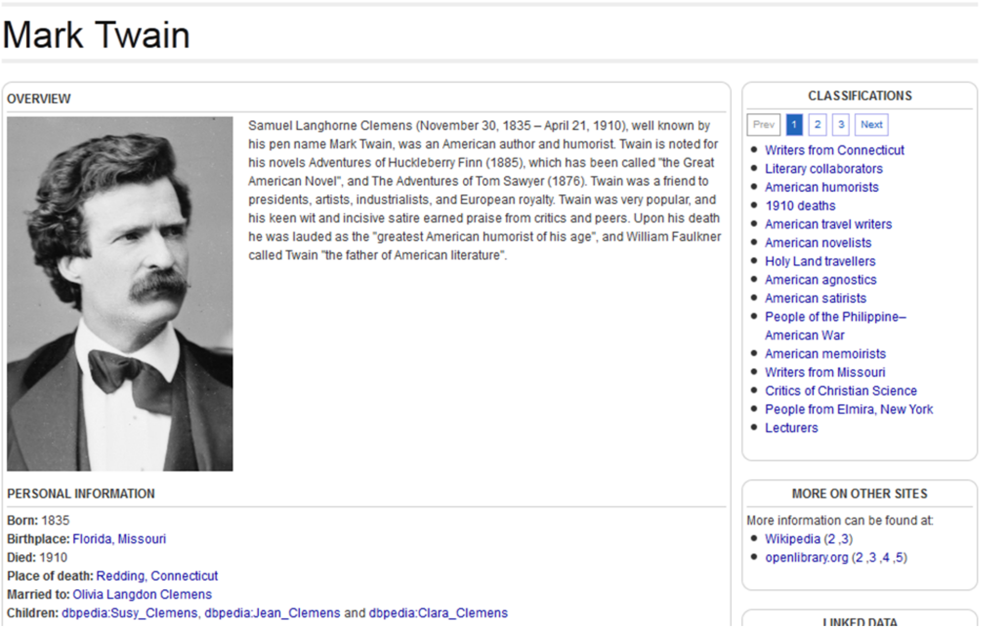
DATASET

Example resources:

- Argentina
- Venice
- London Heathrow Airport
- Chicago
- Mark Twain
- Billy the Kid
- dbpedia:Graceland
- Neil Alden Armstrong
- San Francisco
- groups:421600133093
- gtin:00850167002023
- Civilization
- Office for National Statistics Data
- dbpedia:Rock_music
- authors:OL20157A

[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.



Mark Twain

OVERVIEW

Samuel Langhorne Clemens (November 30, 1835 – April 21, 1910), well known by his pen name Mark Twain, was an American author and humorist. Twain is noted for his novels Adventures of Huckleberry Finn (1885), which has been called "the Great American Novel", and The Adventures of Tom Sawyer (1876). Twain was a friend to presidents, artists, industrialists, and European royalty. Twain was very popular, and his keen wit and incisive satire earned praise from critics and peers. Upon his death he was lauded as the "greatest American humorist of his age", and William Faulkner called Twain "the father of American literature".

PERSONAL INFORMATION

Born: 1835
Birthplace: Florida, Missouri
Died: 1910
Place of death: Redding, Connecticut
Married to: Olivia Langdon Clemens
Children: dbpedia:Susy_Clemens, dbpedia:Jean_Clemens and dbpedia:Clara_Clemens

CLASSIFICATIONS

- Writers from Connecticut
- Literary collaborators
- American humorists
- 1910 deaths
- American travel writers
- American novelists
- Holy Land travellers
- American agnostics
- American satirists
- People of the Philippine–American War
- American memoirists
- Writers from Missouri
- Critics of Christian Science
- People from Elmira, New York
- Lecturers

MORE ON OTHER SITES

More information can be found at:

- Wikipedia (2 ,3)
- openlibrary.org (2 ,3 ,4 ,5)

LINKED DATA

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.

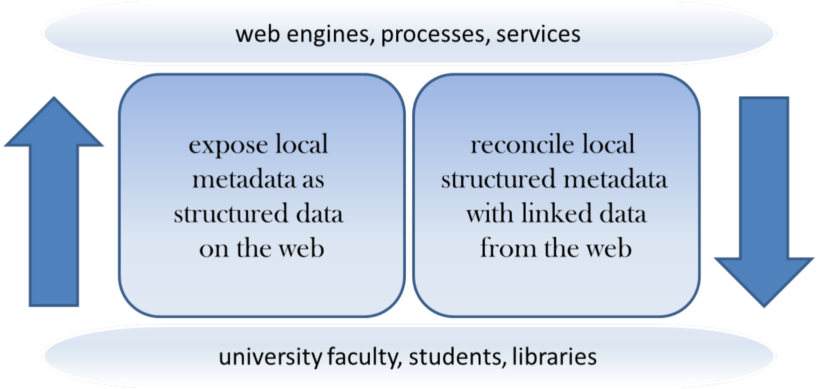
From the library perspective, work will include identifying and gathering

- 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.

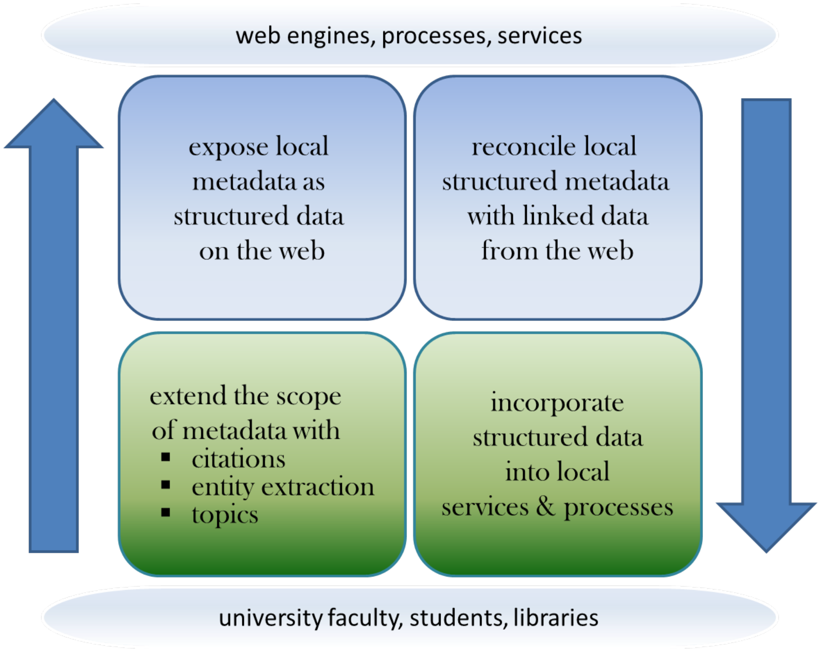
The resulting pool of structured-data statements will feed into an array of web-wide engines, processes and services including projects and services supported by Freebase, Seme4 and <sameAs>, and the Talis Platform (see the section on **Infrastructure** for more details). As seen in the illustration 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.


web engines, processes, services
expose local metadata as structured data on the web
reconcile local structured metadata with linked data from the web
university faculty, students, libraries

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.


web engines, processes, services
expose local metadata as structured data on the web
reconcile local structured metadata with linked data from the web
extend the scope of metadata with
▪ citations
▪ entity extraction
▪ topics
incorporate structured data into local services & processes
university faculty, students, libraries

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;*

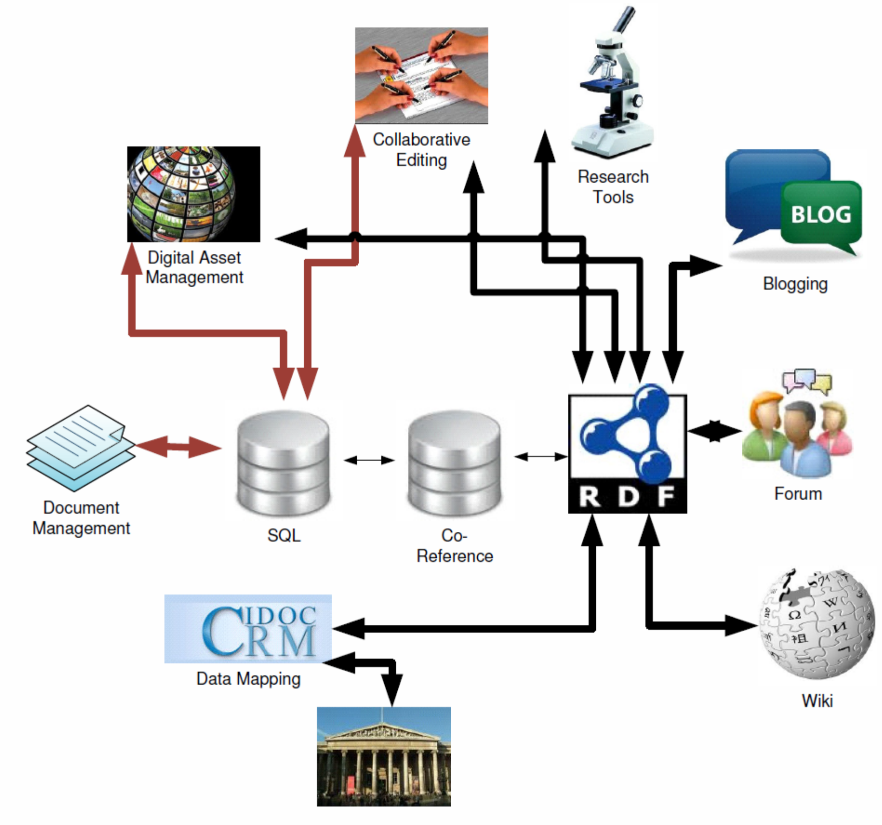
- *Data and digital analysis tools.*
- *Collaboration tools*
- *Semantic RDF data sources*
- *Data and digital management tools.*
- *Internet design and authoring tools*
- *Web Publication*

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 ResarchSpace 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.

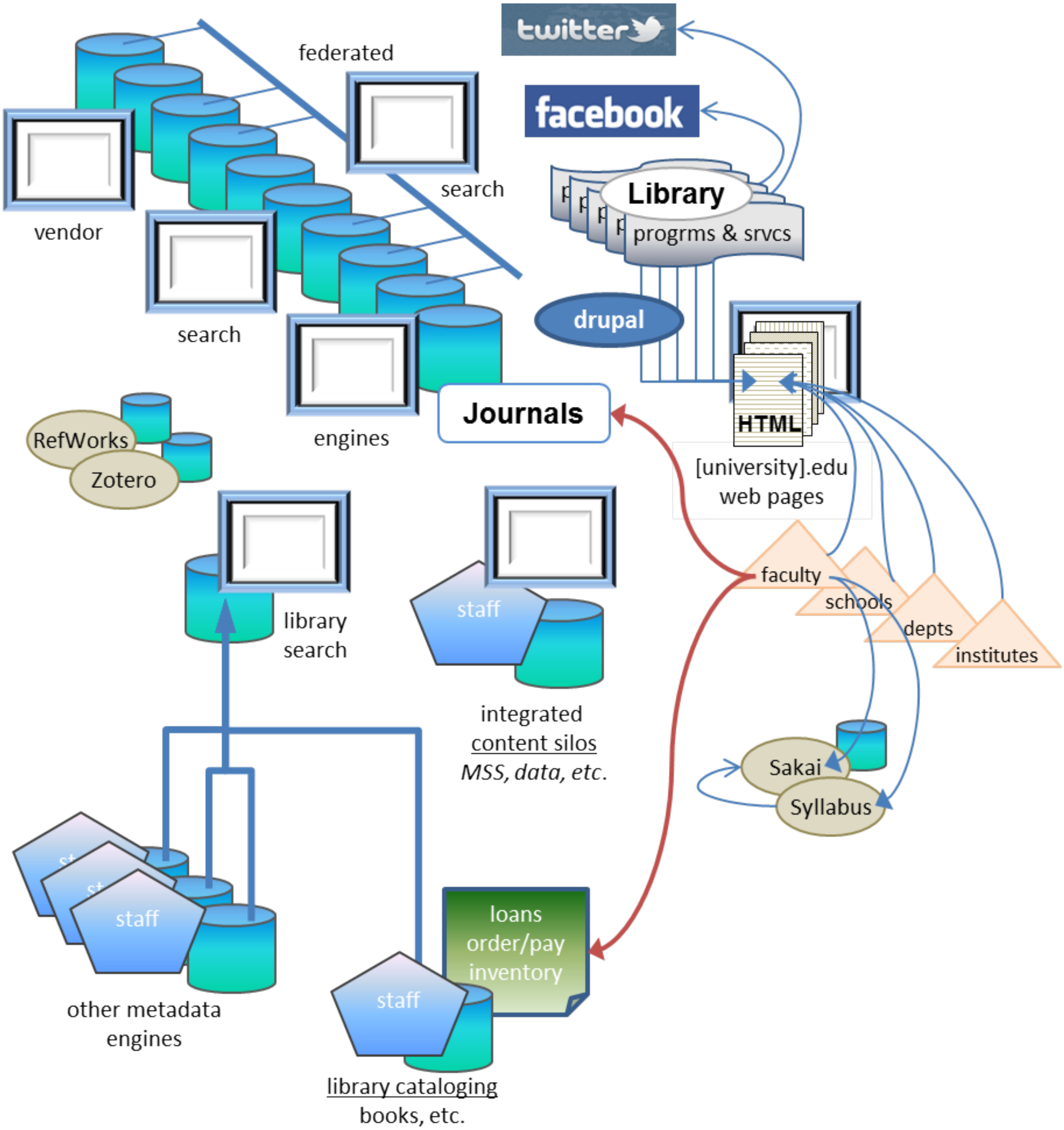
1. Current situation [full-page image]

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



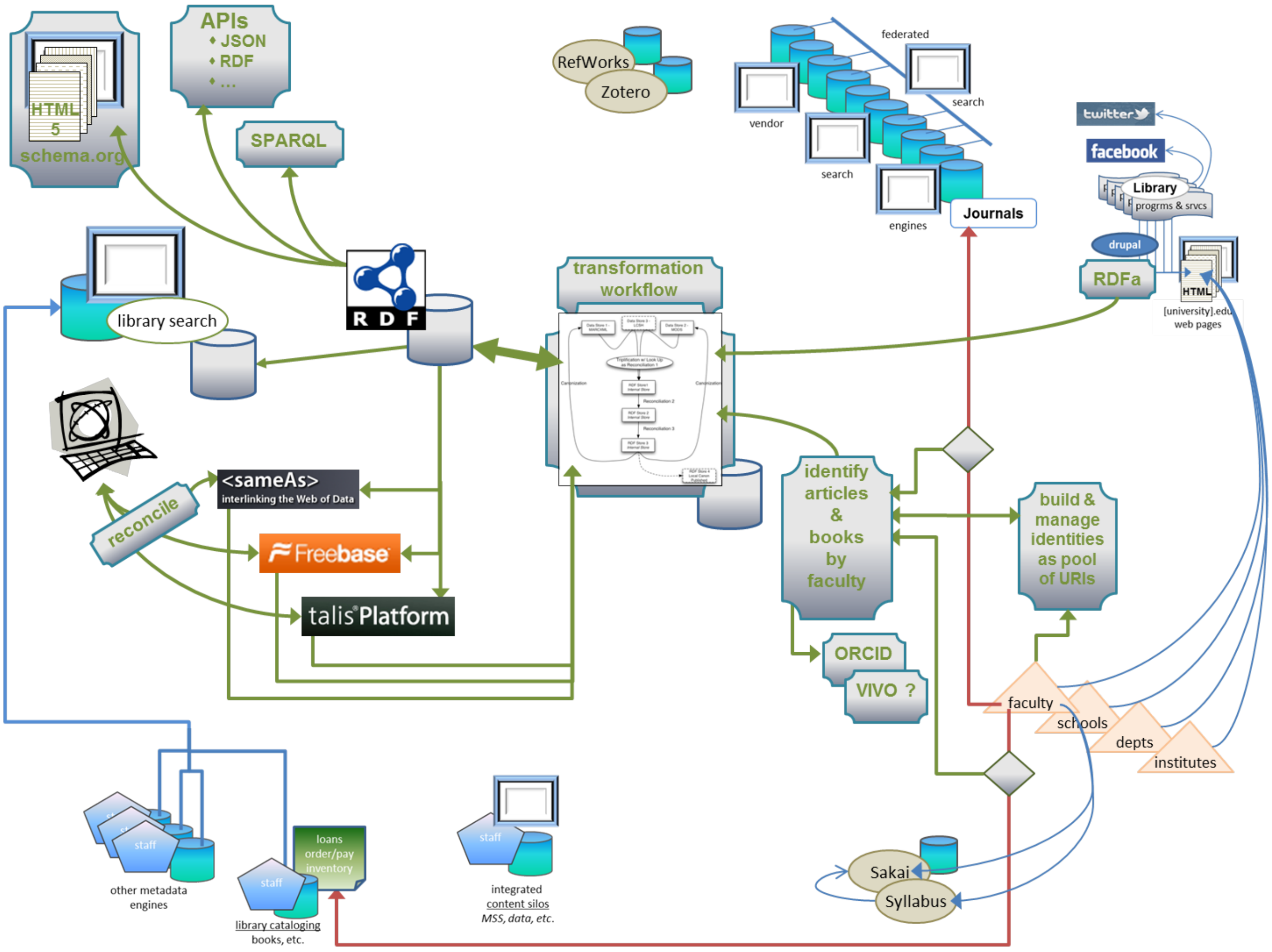
2. Phase-one components

[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



- RDF data store(s) and systems
    - here we plan to draw on the extensive work underway in LOD2 (Linked Open Data 2) projects, work that began in 2010 and is funded at 6+M€ through 2014. Many of the capabilities that this project will require are available or under development in the LOD2 technology stack
    - the aforementioned RFP for the ResearchSpace project will yield further knowledge related to this aspect of the project's environment
    - a third source of knowledge that will help inform this part of the project's planning comes from Talis Consulting through their work with the LATC project (DERI, Galway plus Talis Consulting and others). Details are here.
- schema.org (HTML5), APIs (JSON, RDF), SPARQL, library search
    - exposing linked-data in ways that allow it to flow out into the web-wide pool of structured data includes these (and other) types of processes and tools
    - the overall objective here is publishing the project's linked-data so that other like-minded efforts can make use of the project's work
    - down the road, open-access to the data will feed a self-sustaining create-use-curate-extend cycle through which structured data *about* content will grow and evolve to include links that describe *how content is used* by mining the context of courses, the context of citations embedded in articles and books, mining the context of the products of day-in-and-day-out research work via tools like RefWorks and Zotero
    - schema.org is the search-engine community's current quickly evolving approach to embedding structured data in all manner of web pages:
        - summarized in the Survey developed for the Stanford Workshop
        - included in phase one as part of end-to-end data flows for the project
    - APIs include interactive capabilities for extracting linked data from the local pool of RDF, as well as scheduled exports of the entire structured data store
    - SPARQL is the linked-data equivalent of SQL for relational databases, providing an interactive query language for exploring linked-data stores
- <sameAs>, Freebase, Talis Platform
    - a key component of the transformation workflow is work commonly known as reconciliation—processes (machine and human) by which one makes statements about the relationship between two URIs, *e.g.*, these two URIs refer to the person known as Mark Twain (in linked-data terms, this URI is the sameAs that URI)
    - including these three services in the workflow recognizes them (and others that will emerge) as environments in which ongoing processes (machine and human) work to identify the relationships between entities (linked-data URIs)
    - the scope of these processes includes access to web-wide pools of links
    - the flow of data back into the project's *transformation workflow* denotes our interest in taking advantage of extant service and processes to help refine reconciliation within the local store of linked data
- with the exception of adding phase-one linked-data facts to the pool of data indexed by the library search engine, other components of the environment remain as they were in the present-day sketch of components

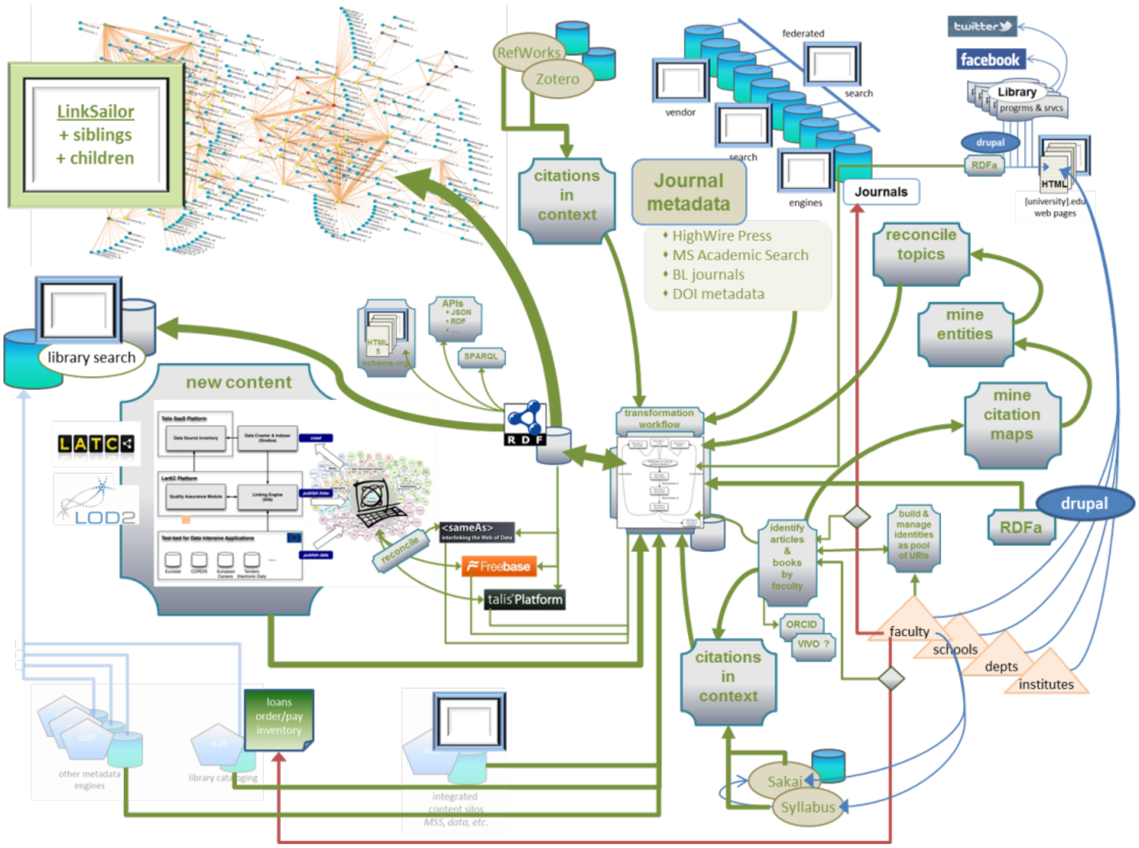
3. Phase-two components

[full-page image]

Moving clockwise around the sketch, we see

- faculty, schools, departments, institutes, etc.
    - further in-depth analysis of the pool of faculty books, articles, data repositories etc.
    - here begins the work of mining content vehicles for facts that go beyond the people, places, organizations, events, etc. associated with publications
    - aimed at uncovering the context of relationships between published artifacts, efforts will include:
        - mining the citation maps embedded in publications and feeding those links into the university-wide pool of structured data
        - mining entities found inside the text publications, a process that extends the depth of descriptive facts associated with a work
        - using the results of entity mining to extend the breadth and scope of topical analysis of content
    - reconciling topics will undertake the first phases of crosslinking vocabularies and taxonomies in order to bridge between varied pools of topical terminologies and strucures, *e.g.,* Library of Congress Subject Headings, the taxonomies and vocabularies embedded in journal environments (HighWire Press is one example), and topic cross-linking being investigated in Mike Bergman's work on UMBEL
    - note the Drupal←→RDFa addition to flow of information into academic web sites, this to denote increased use of linked-data as an increasingly common component of web environments built and managed with course/learning management engines
- course management and related tools
    - in the Stanford case, these include Sakai and a service dubbed Syllabus which gathers reading lists like information about a substantial number of courses each quarter
    - here begins the work of mining tools and services for facts that denote how faculty and their courses make use of various types of content

Phase two components.



- library acquisitions, cataloging, circulation, and inventory management
  + other metadata engines
  + vertically integrated content silos
  + library search
    - phase two of the project greys-out most aspects of the library environment
    - this to denote the role of these systems moving to legacy status—a status having decreasing levels of maintenance for the metadata they support
      - one can expect circulation, acquisitions & inventory management to be the longest lasting component of legacy systems
      - the access front-ends for vertically integrated content silos are another component that will be around for some time
    - the greyed-out lines connected to library search denote continued use of non-LMS search engines as predominant form of key-work searching
    - the heavier green lines from metadata stores for legacy system represent migration of all local metadata pools into the structured data workflow
- new content (LATC, LOD2)
    - here begins the work of creating structured data for new resources coming into a research library's services and programs
      - references to LATC and LOD2 point out the increasing level of work across the linked-data community on the tools required to build and manage structured data as a production level set of workflows and processes
      - other projects are looking at various aspects of these types of engines, e.g., eXtensible Catalog, University of Rochester) and ANDS (Australian National Data Service)
      - note that the pool of structured data on which such production environments will be based is in fact the very same set of curated reconciled links that are being developed by services like <sameAs>, Freebase, Talis, and others.
- LinkSailor +siblings +children
    - here begins the use of emerging tools that will support navigation of structured-data graphs as an alternative to keyword, faceted search engines
    - this environment will be characterized by a long-lived evolution, starting with the relatively simple but extensible features inherent in LinkSailor and similar efforts
    - during that evolution, keyword-faceted search engines will benefit a great deal from the increased breadth and depth of metadata to be supplied from pools of linked data
- RefWorks/Zotero
    - here begins another facet of the work of mining tools and services for facts that denote how all members of the academic community make use of various types of content—facts about the contexts in which content is used
- Journal metadata
    - the work of identifying articles by current members of the local faculty will involve taking a look at metadata for the content of journals in a new way

  - the primary focus will be on sources of information related to authors' affiliations when the article was written
  - working though available and emerging resources may well provide an opportunity to expand the amount of metadata that could be included in a local pool of structured data, for example:
    - in the Stanford case, one can expect to see full schema.org treatment of articles for HighWire publishers and for all of PubMed's abstracts
    - the Workshop at Stanford revealed that the British Library has a large, cross-discipline collection of article-level metadata for ca. 20,000 journals with publication dates that span more than a decade.
    - there are indications that the full range of metadata associated with CrossRef's DOIs is coming into the public domain.

Materials & URIs

Materials

On the **explicit metadata** side of the equation, we will include the content bearing formats that most commonly appear as the published products of the faculty's research and scholarship. This will confine the project within manageable boundaries, letting us focus our work on the same range of materials for each of the intended dual focal points (the academy and the library).

For the academy side of the equation, the focus will be on identifying the published work of Stanford faculty and graduate students. This in order to meet the project's goal:

> *allow an academic institution, its faculty and students, and its libraries to operate as full-fledged participants and 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.*

We will address the library's complete pools of metadata for those types of resources being created by academic side of the house. Taken together this array of materials will include:

- articles
  - STM
    - HighWire Press[8] – articles and PubMed abstracts
    - pursue other sources of affiliation data
  - humanities & social sciences[9]
    - HighWire Press – articles (e.g. Oxford University Press and Sage)
    - pursue cooperation with Microsoft's Acadmic Search project
    - pursue other sources of affiliation data

[8] HighWire Press is committed to emitting Linked Data for all the articles streaming through its services.

[9] At the Stanford Linked Data Workshop an intriguing possibility surfaced for transcoding to Linked Data the metadata from 20,000 journals for which the British Library has the rights to manipulate the metadata. This we intend to pursue with our colleagues at the British Library in a separate project.

- string matching with university faculty names
    - DOIs – in the process of going public as linked data
    - schema.org level data: name, title, journal, date, vol, pgs
    - pursue use of arXiv data
    - pursue use of the BL journal citation data

- books, including dissertations and theses
    - take advantage of cataloging for Stanford authors
    - use string matching in OCLC searches
    - all books published by Stanford University Press and other Stanford publishing enterprises (e.g. CSLI publications, Hoover Institution).

With respect to resources that embody **implicit forms of metadata**, the project will strive (and limit itself) to retrospectively re-engineering locally produced products scholarship, research, and teaching. This in order to:

- capture the academy's expert opinions about relationships between and among discrete pieces of published content
- distill those connections into structured data that expresses those correlations as highly-refined, navigable links amongst resources in the web of data.

We expect such information to become an extremely valuable component of linked data, one that refines that quality and extends the reach of structured data that is derived from traditional forms of factual and topical metadata. This effort will be focused on facts that go beyond describing individual publications. We can gain access through this sub-project to the context of how knowledge and information resources are used throughout the work of all members of the academic and library communities.

On the other hand, current technology and workflows do little to provide functional access to this type of information. An exception to this state of affairs are the citations of related works that are available in journal articles that are published on the web via services like HighWire Press.

Given the difficulty of dealing with this type of data and content in its present state, the project would limit its pursuit of means and methods to exploit implicit metadata to the resources created by members of each institution implementing this plan. This limit will produce a viable prototype for this type of effort, while at the same time limiting the scope of work to manageable proportions. Resources will include:

1. course and teaching content in the Stanford case
    - CourseWork (a local instantiation of Sakai) -- this will involve reverse engineering some of its modules
    - Syllabus – substantive use on campus, will require reverse engineering
    - stanford.edu – pursue whether to mine stanford.edu for course sites

2. citations found within publications [use Stanford authorship as a lens]
    - journals
        - HighWire – citation maps exist

        - pursue cooperation with Microsoft's Acadmic Search project
        - stanford.edu – possibly mine stanford.edu for open source copies
        - pursue other sources of embedded citation data
    - books and dissertations
        - mine Google scans of Stanford authored works for citations
          -- send thru any faculty authored books not already scanned
    - consider using commercial entity and structure mining engine(s) for Stanford authored articles and other publications
        - pursue in order build substantive, effective prototype for the value of this type of linked data

3. bibliographic and citation management tools
    - Zotero – will require considerable work with the community
        - what is the incentive to share, what will they gain by sharing?
    - RefWorks is other tool with substantive amounts of use on campus

4. Stanford University Press and Hoover Institution books

Other institutions implementing the model will likely make analogous selections reflecting their particular situations.

URIs for people, organizations, publications …

1. linked-data statements for members of the academic community
    a. people
    - sources of data in the Stanford case
        - at Stanford, make use of emerging CAP (Community Academic Profiles) effort [more information here]. Currently some 4,000 profiles of School of Medicine's faculty, academic staff, postdocs, and students. The environment generates basic profiles that give an integrated portrait of each person's activities—University appointments/affiliations, research interests and publications. People have control over the content including text for describing research interests, ability to control which publications are included, uploads for CV and photos, along with provision for other types of information including contact information, awards and honors, community and global work, etc.
        - take advantage of other journal/report author name authority projects such as ORCID, MIMAS, ISNI, VIVO
        - NB: projects will need to address opt in/out considerations & FERPA

    b. organization
    - schema and data collection for modeling academic organizations
        - look especially at what Southampton has done
            - Oct 2010
            - Mar 2011
    - note that the schema for campus ID data includes affiliation flags

c. work with and contribute to projects focusing on academic people

- ORCID Open Researcher & Contributor ID
- ISNI International Standard Name Identifier
- JISC Names Project 2011 status report

d. projects/tools that mix mapping/mining people, organizational structure, and content

- o Open University
    - ▪ especially the LUCERO project
- o VIVO
    - ▪ "national network across all science disciplines"

3. Tools and resources for library content

a. name authorities

- Library of Congress name authority files are freely available for use now
    - o pursue "authority record" for every name (as DNB does for Germany)
    - o British Library authorities included with US data
    - o pursue access to DNB and BnF data
- VIAF
    - o pursue means for making open (or at least CC BY) use of this

b. topics

- Map the upper levels of HighWire Press' subject taxonomy to equivalents in the Library of Congress Subject Headings to create topic links between book and article literature
- Mike Bergman's thinking and work related to UMBEL warrants consideration
- Capture additional names, organizations, and events through ell-established entity extraction techniques
- Assign additional controlled subject terms at the chapter level through semantic analysis technologies (e.g. TEMIS)

c. other types of useful data

- Incorporate a wealth of publisher data stored in ONIX files for individual publications

4. Making and reconciling linked-data statements for academic & research library content

a. data model as first, essential design task

- project intends to pursue a model that can, at minimum: represent all the materials slated for inclusion at level comparable to schema.org
- pursue needed enhancements of schema.org per HighWires' successes
- consult with BL, Talis, Freebase, and Google's schema.org staff
- Pursue collaboration with the Program for Cooperative Cataloging as it develops a model of the essential elements of bibliographic data

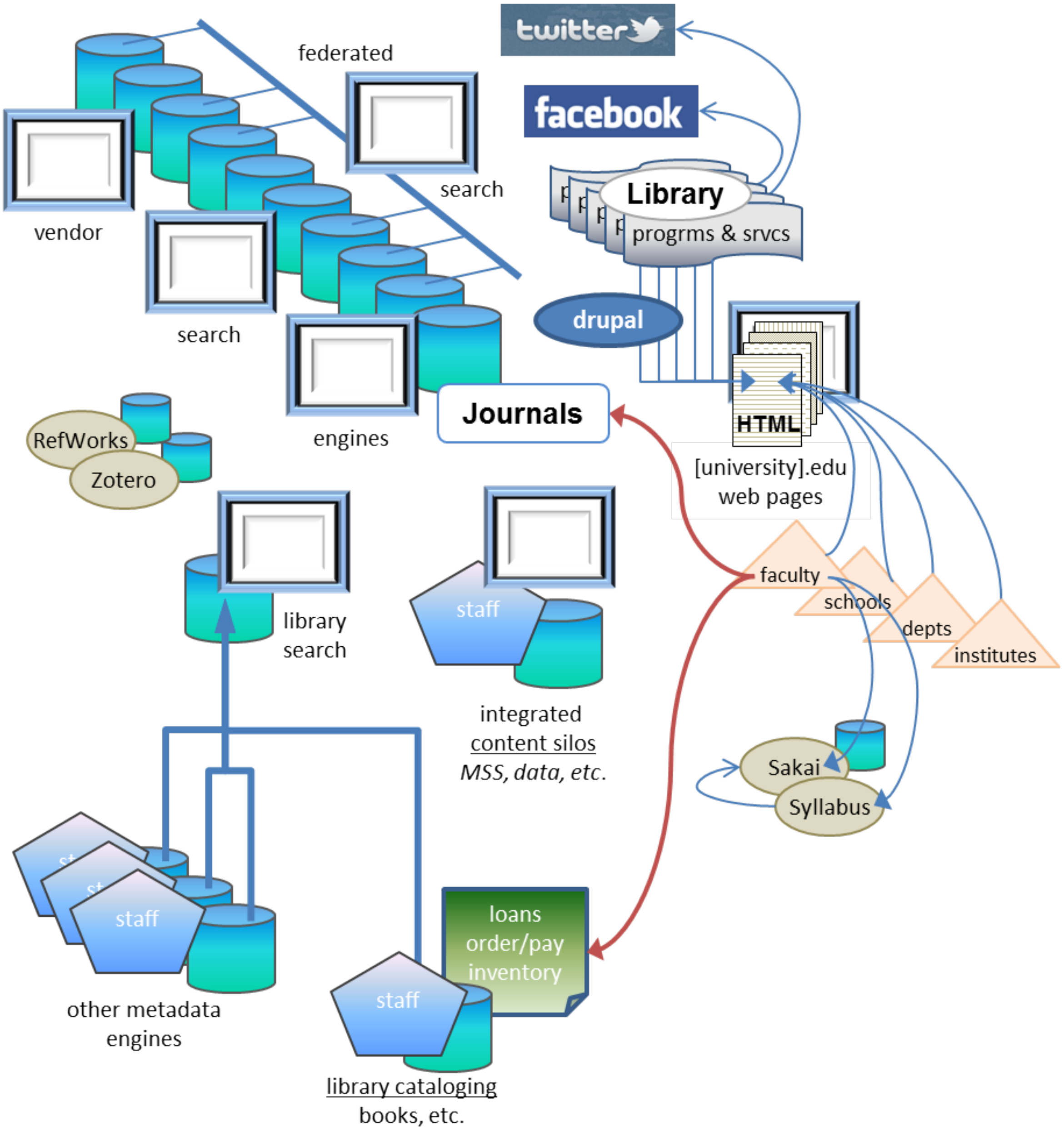
- Partner with the eXtensible Catalog project building and elaborating utilities to create an open source engine for transforming MARC data to RDF triples conforming to the project's data model

Potential Partnerships

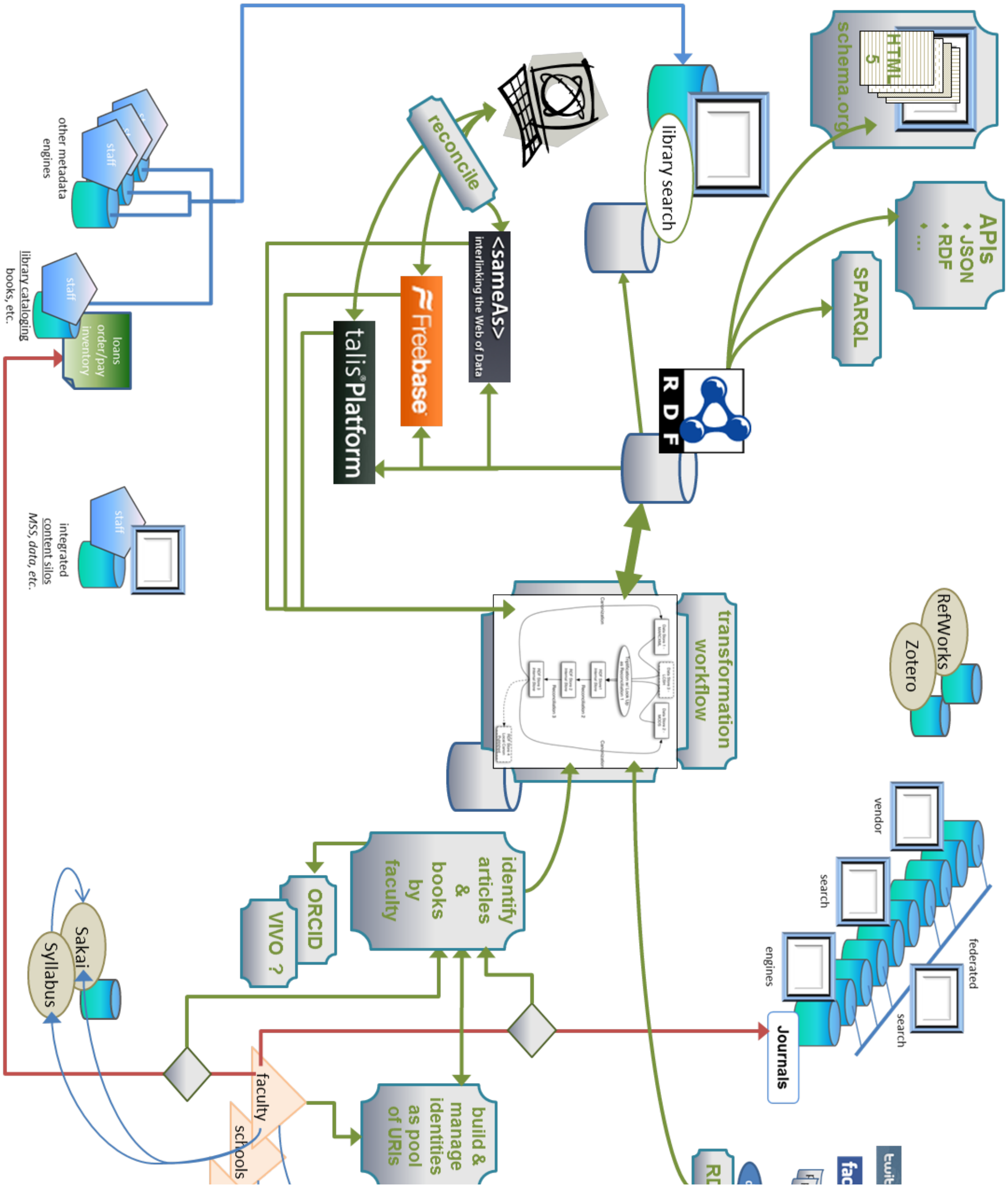
1. Talis
   - Talis Consulting
     - strategic planning
     - development of data model (they consulted with BL)
   - Talis Platform and LATC Interlinking Platform
     - exposure of project's linked data products in web-wide venue
     - reconciliation of URIs to enhance local pool of data
     - development of linked-data creation/mgmt. tools, process, etc.
2. British Library
   - consulting with Metadata Services on model development
   - pursue access to, use of, and development of journal metadata from 20,000 journals
3. ResearchSpace
   - make use of their RFP
   - consult with the project's technical design team
4. European national libraries
   - pursue working with the 46 member countries of the CENL in developing an international authority file based on their individual national authority files; select European libraries have emitted some of their bibliographic metadata as Linked Data (British Library), some who anticipate doing so soon (Deutsche Nationalbibliothek), some of whom might be persuaded to let Stanford transcode the data (perhaps, tentatively the Bibliothéque nationale de France); We would add linked data resources emitted by the Scandinavian libraries (Sweden, Norway) that have done so and the work of the Finns from Aalto University
5. Microsoft Academic Search
   - Initial conversations with directors of Microsoft Academic Search indicate strong interest in transcoding metadata they are currently using in Microsoft Academic Search and emitting RDF triples/URIs to open stores
6. Freebase
   - exposure of project's linked data products in web-wide venue
   - reconciliation of URIs to enhance local pool of data
   - consultation regarding data model & development of processes/ workflows
7. Seme4 and <sameAs>
   - exposure of project's linked data products in web-wide venue
   - reconciliation of URIs to enhance local pool of data
   - consultation regarding data model & development of processes/workflows
8. JISC
   - liaison to like projects and efforts in the UK
9. ANDS (Australian National Data Service) liaison to like projects and efforts in Australia

Full page images

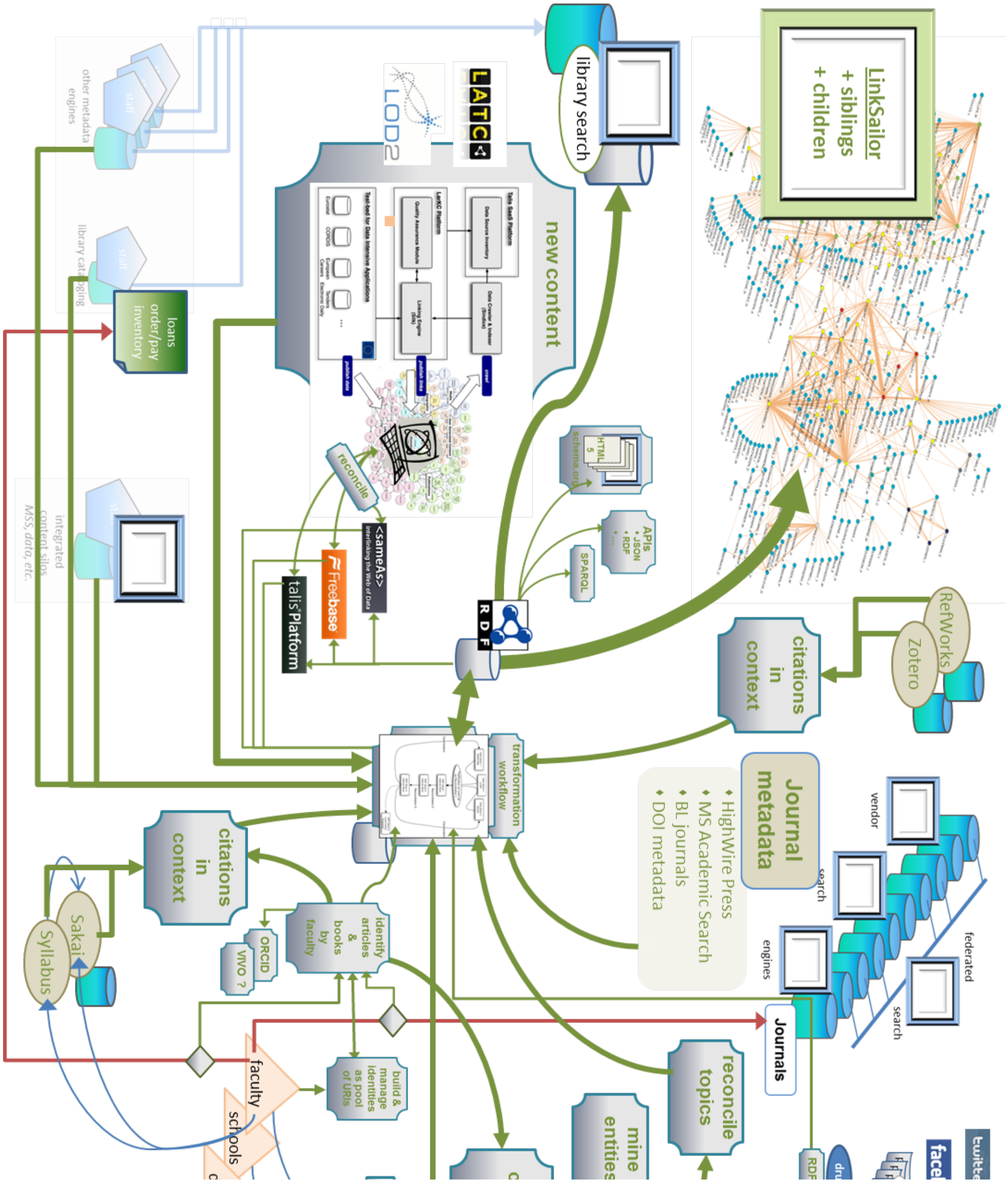
1. Schematic snapshot – current characteristics



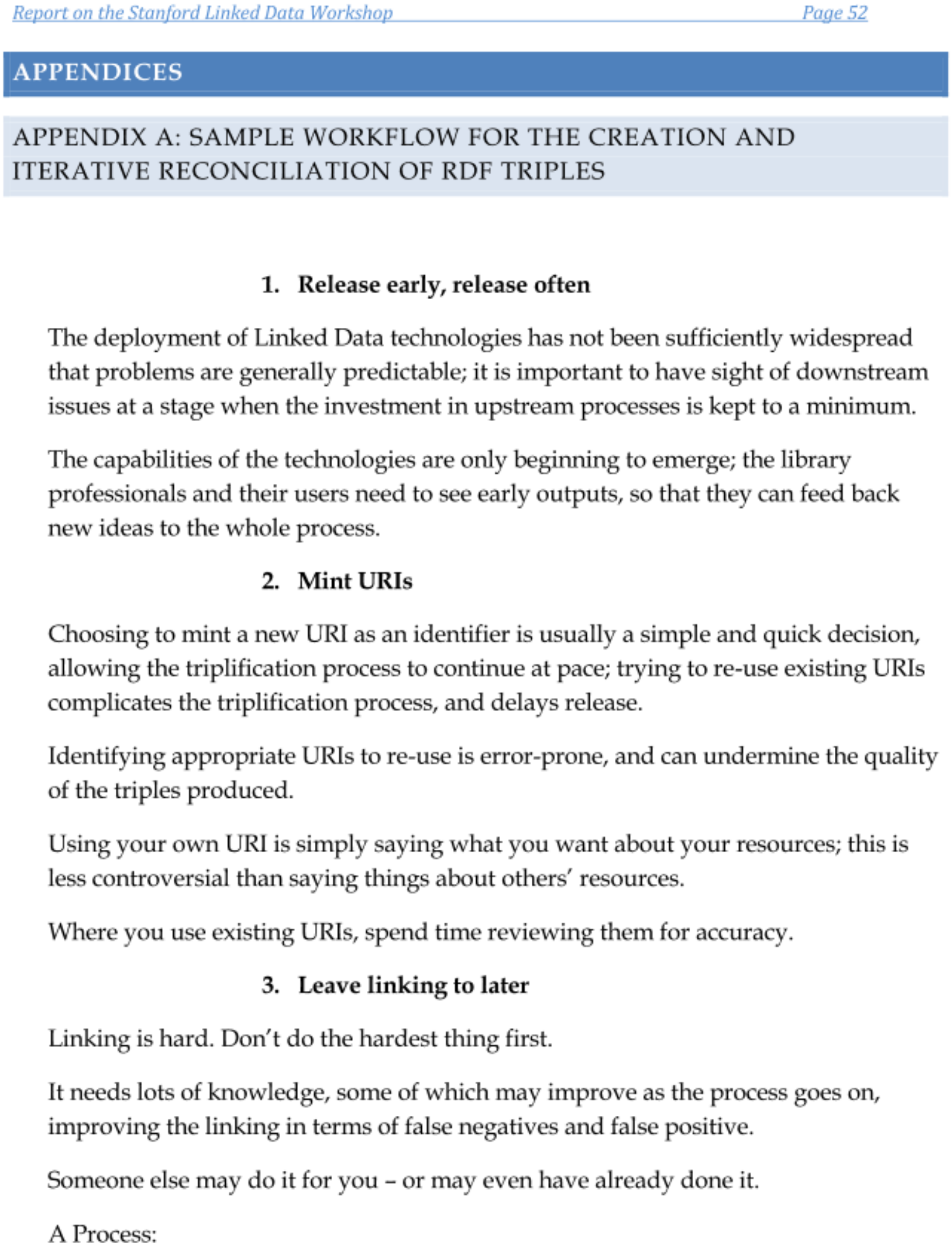
2. Schematic snapshot – phase-one components



3. Schematic snapshot – phase-two components



4. Transformation Workflow from Linked Data Workshop at Stanford:



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

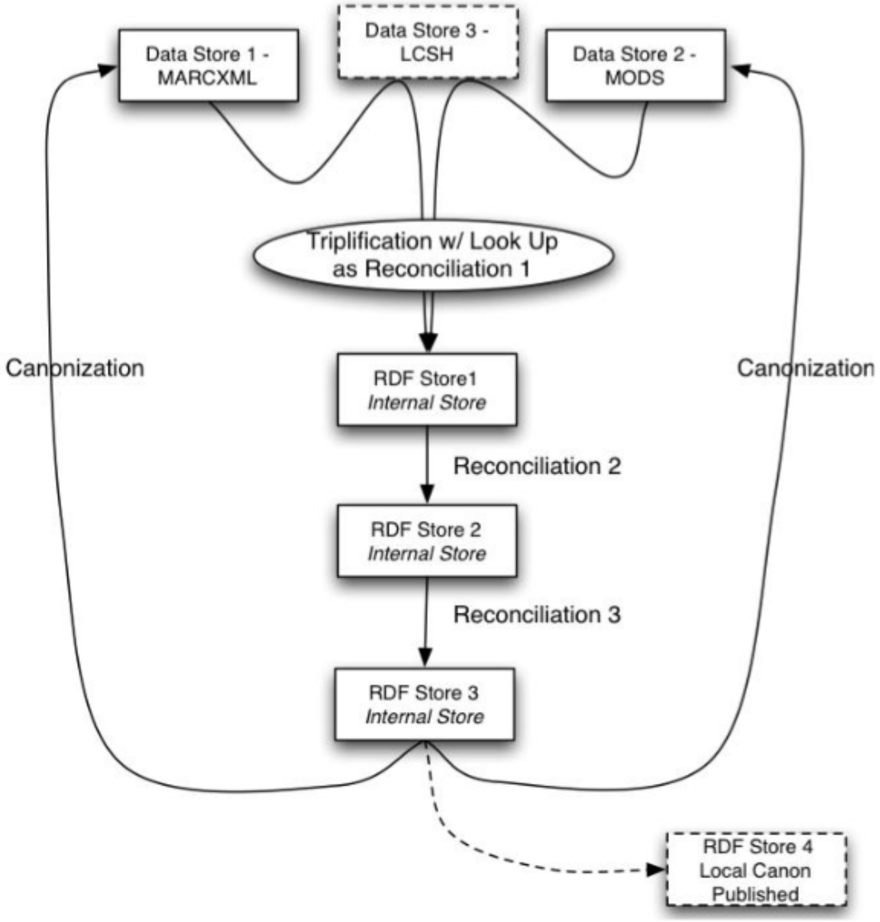
Linking is hard. Don't do the hardest thing first.

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.

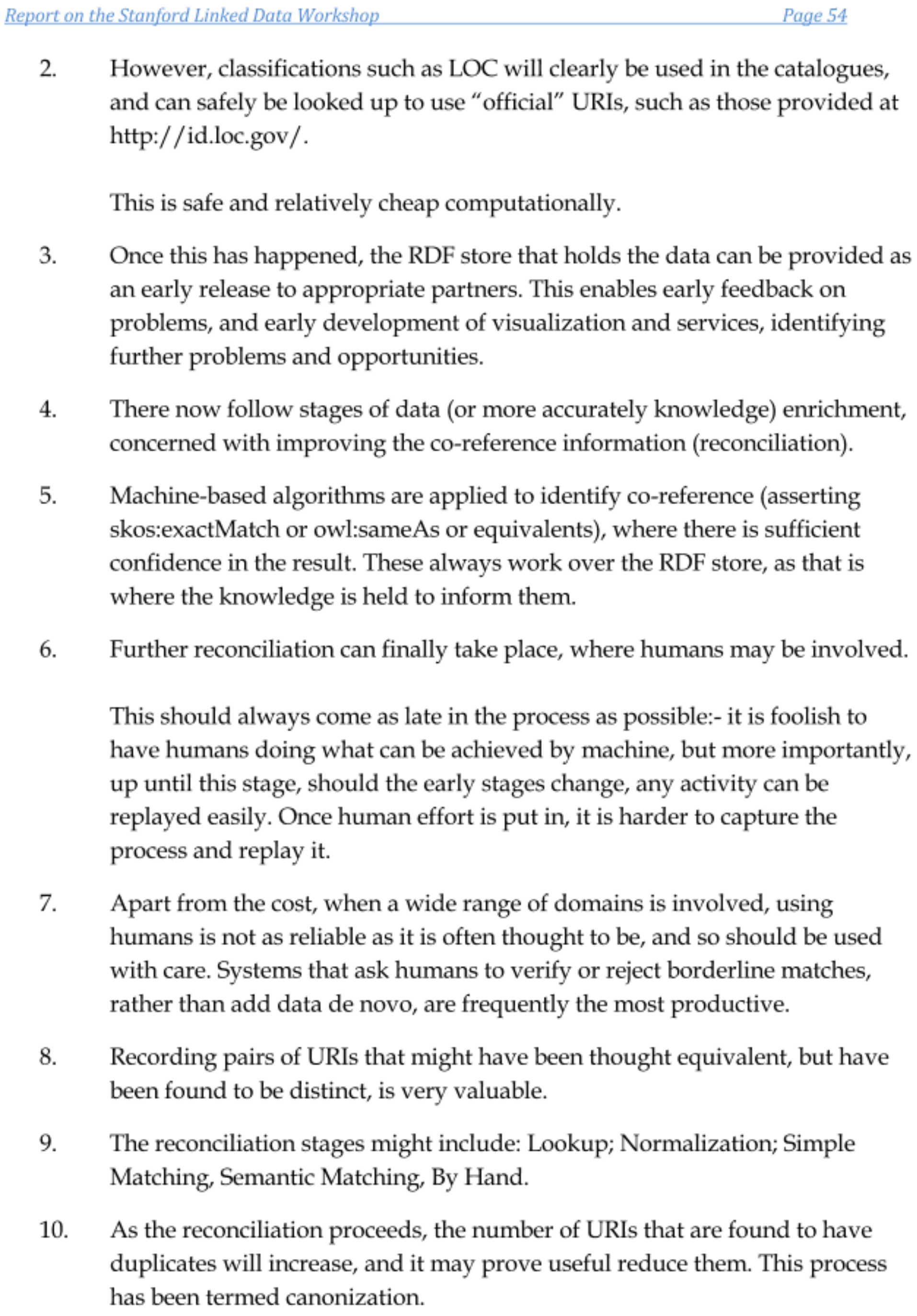
A Process:

Report on the Stanford Linked Data Workshop Page 53


Data Store 1 - MARCXML
Data Store 3 - LCSH
Data Store 2 - MODS
Triplification w/ Look Up as Reconciliation 1
Canonization
Canonization
RDF Store1 Internal Store
Reconciliation 2
RDF Store 2 Internal Store
Reconciliation 3
RDF Store 3 Internal Store
RDF Store 4 Local Canon Published

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.

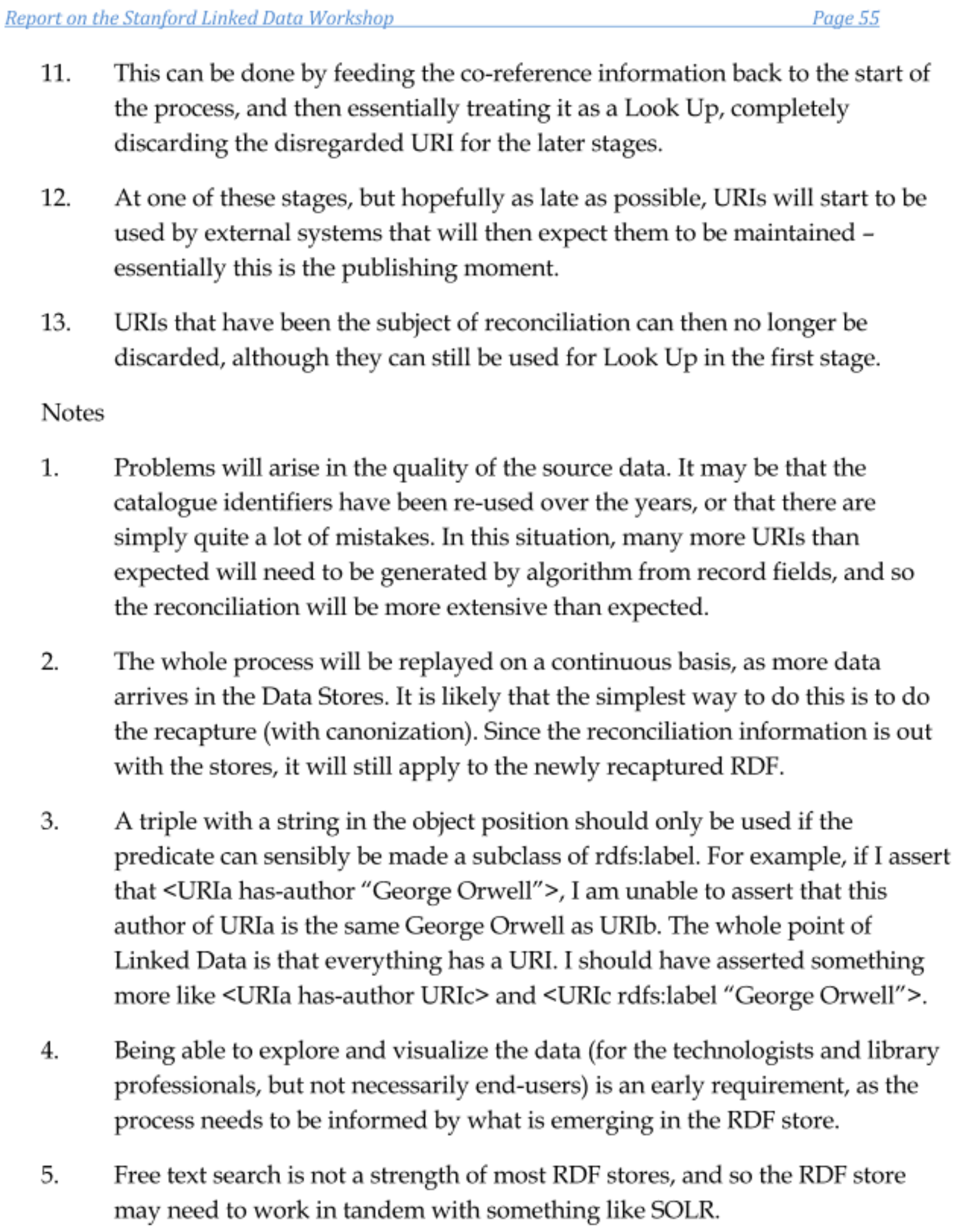
   This should always come as late in the process as possible:- it is foolish to have humans doing what can be achieved by machine, but more importantly, up until this stage, should the early stages change, any activity can be replayed easily. Once human effort is put in, it is harder to capture the process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

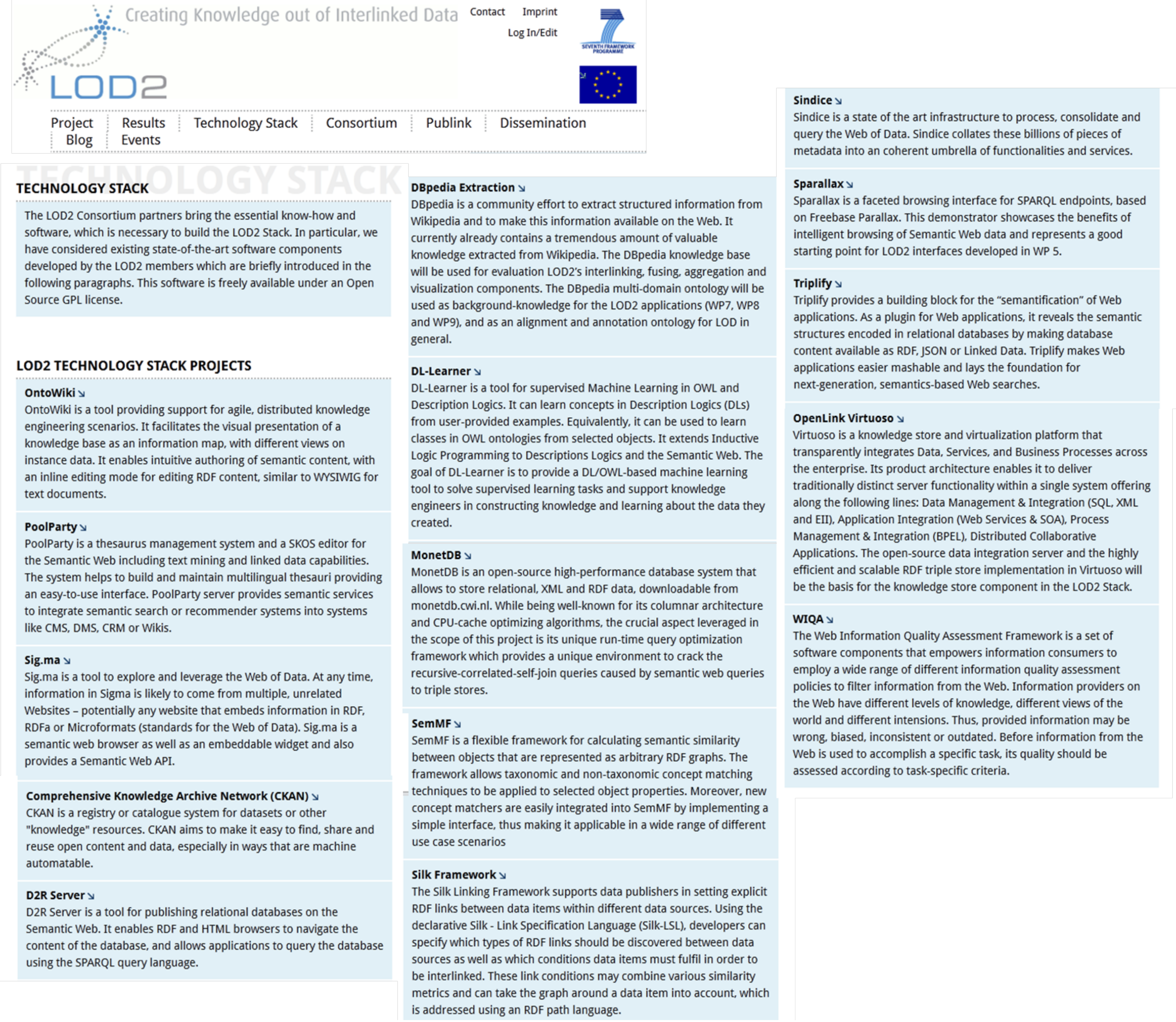
2. The whole process will be replayed on a continuous basis, as more data arrives in the Data Stores. It is likely that the simplest way to do this is to do the recapture (with canonization). Since the reconciliation information is out with the stores, it will still apply to the newly recaptured RDF.

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.

5. LOD2 (Linked Open Data 2 **)** Technology Stack, Summary, Home



TECHNOLOGY STACK

The LOD2 Consortium partners bring the essential know-how and software, which is necessary to build the LOD2 Stack. In particular, we have considered existing state-of-the-art software components developed by the LOD2 members which are briefly introduced in the following paragraphs. This software is freely available under an Open Source GPL license.

LOD2 TECHNOLOGY STACK PROJECTS

OntoWiki ↘
OntoWiki is a tool providing support for agile, distributed knowledge engineering scenarios. It facilitates the visual presentation of a knowledge base as an information map, with different views on instance data. It enables intuitive authoring of semantic content, with an inline editing mode for editing RDF content, similar to WYSIWIG for text documents.

PoolParty ↘
PoolParty is a thesaurus management system and a SKOS editor for the Semantic Web including text mining and linked data capabilities. The system helps to build and maintain multilingual thesauri providing an easy-to-use interface. PoolParty server provides semantic services to integrate semantic search or recommender systems into systems like CMS, DMS, CRM or Wikis.

Sig.ma ↘
Sig.ma is a tool to explore and leverage the Web of Data. At any time, information in Sigma is likely to come from multiple, unrelated Websites – potentially any website that embeds information in RDF, RDFa or Microformats (standards for the Web of Data). Sig.ma is a semantic web browser as well as an embeddable widget and also provides a Semantic Web API.

Comprehensive Knowledge Archive Network (CKAN) ↘
CKAN is a registry or catalogue system for datasets or other "knowledge" resources. CKAN aims to make it easy to find, share and reuse open content and data, especially in ways that are machine automatable.

D2R Server ↘
D2R Server is a tool for publishing relational databases on the Semantic Web. It enables RDF and HTML browsers to navigate the content of the database, and allows applications to query the database using the SPARQL query language.

DBpedia Extraction ↘
DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. It currently already contains a tremendous amount of valuable knowledge extracted from Wikipedia. The DBpedia knowledge base will be used for evaluation LOD2's interlinking, fusing, aggregation and visualization components. The DBpedia multi-domain ontology will be used as background-knowledge for the LOD2 applications (WP7, WP8 and WP9), and as an alignment and annotation ontology for LOD in general.

DL-Learner ↘
DL-Learner is a tool for supervised Machine Learning in OWL and Description Logics. It can learn concepts in Description Logics (DLs) from user-provided examples. Equivalently, it can be used to learn classes in OWL ontologies from selected objects. It extends Inductive Logic Programming to Descriptions Logics and the Semantic Web. The goal of DL-Learner is to provide a DL/OWL-based machine learning tool to solve supervised learning tasks and support knowledge engineers in constructing knowledge and learning about the data they created.

MonetDB ↘
MonetDB is an open-source high-performance database system that allows to store relational, XML and RDF data, downloadable from monetdb.cwi.nl. While being well-known for its columnar architecture and CPU-cache optimizing algorithms, the crucial aspect leveraged in the scope of this project is its unique run-time query optimization framework which provides a unique environment to crack the recursive-correlated-self-join queries caused by semantic web queries to triple stores.

SemMF ↘
SemMF is a flexible framework for calculating semantic similarity between objects that are represented as arbitrary RDF graphs. The framework allows taxonomic and non-taxonomic concept matching techniques to be applied to selected object properties. Moreover, new concept matchers are easily integrated into SemMF by implementing a simple interface, thus making it applicable in a wide range of different use case scenarios

Silk Framework ↘
The Silk Linking Framework supports data publishers in setting explicit RDF links between data items within different data sources. Using the declarative Silk - Link Specification Language (Silk-LSL), developers can specify which types of RDF links should be discovered between data sources as well as which conditions data items must fulfil in order to be interlinked. These link conditions may combine various similarity metrics and can take the graph around a data item into account, which is addressed using an RDF path language.

Sindice ↘
Sindice is a state of the art infrastructure to process, consolidate and query the Web of Data. Sindice collates these billions of pieces of metadata into an coherent umbrella of functionalities and services.

Sparallax ↘
Sparallax is a faceted browsing interface for SPARQL endpoints, based on Freebase Parallax. This demonstrator showcases the benefits of intelligent browsing of Semantic Web data and represents a good starting point for LOD2 interfaces developed in WP 5.

Triplify ↘
Triplify provides a building block for the "semantification" of Web applications. As a plugin for Web applications, it reveals the semantic structures encoded in relational databases by making database content available as RDF, JSON or Linked Data. Triplify makes Web applications easier mashable and lays the foundation for next-generation, semantics-based Web searches.

OpenLink Virtuoso ↘
Virtuoso is a knowledge store and virtualization platform that transparently integrates Data, Services, and Business Processes across the enterprise. Its product architecture enables it to deliver traditionally distinct server functionality within a single system offering along the following lines: Data Management & Integration (SQL, XML and EII), Application Integration (Web Services & SOA), Process Management & Integration (BPEL), Distributed Collaborative Applications. The open-source data integration server and the highly efficient and scalable RDF triple store implementation in Virtuoso will be the basis for the knowledge store component in the LOD2 Stack.

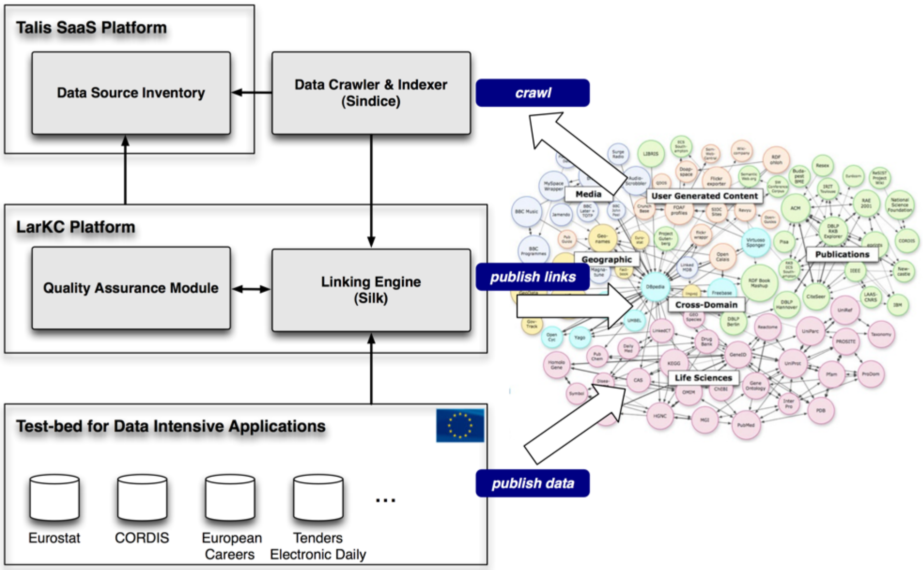
WIQA ↘
The Web Information Quality Assessment Framework is a set of software components that empowers information consumers to employ a wide range of different information quality assessment policies to filter information from the Web. Information providers on the Web have different levels of knowledge, different views of the world and different intensions. Thus, provided information may be wrong, biased, inconsistent or outdated. Before information from the Web is used to accomplish a specific task, its quality should be assessed according to task-specific criteria.

6. LATC (Linked data Around The Clock) -- DERI at Galway, Talis Consulting, and others

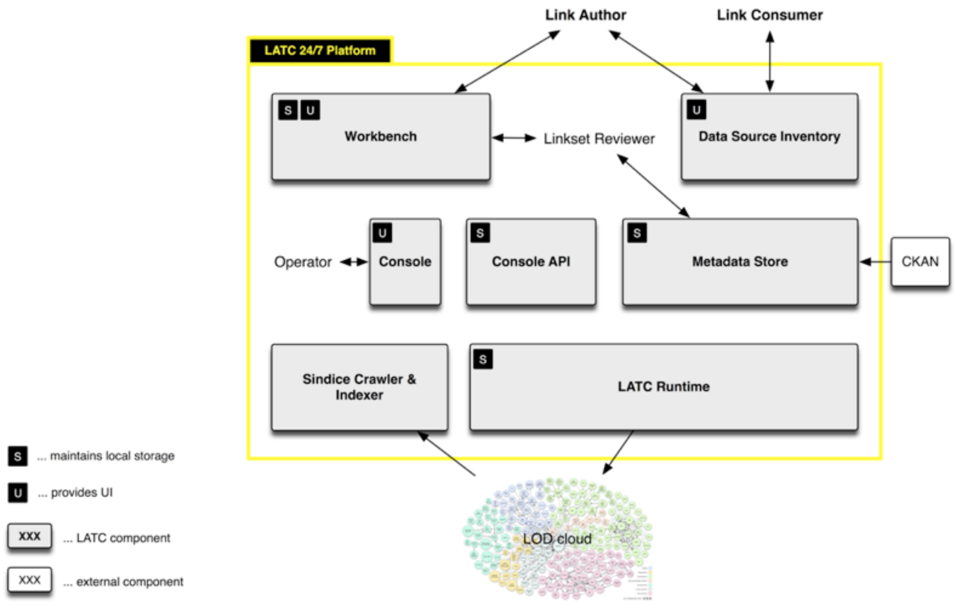
LATC about the project

Putting the Links into Linked Data (a September 2011 post from Talis) provides a description of the tools being incorporated in this project's *Linking Platform*.

architecture sketch:



another view:


Link Author
Link Consumer
LATC 24/7 Platform
S U
Workbench
Linkset Reviewer
U
Data Source Inventory
Operator
U
Console
S
Console API
S
Metadata Store
CKAN
Sindice Crawler & Indexer
S
LATC Runtime
LOD cloud
S ... maintains local storage
U ... provides UI
XXX ... LATC component
XXX ... external component