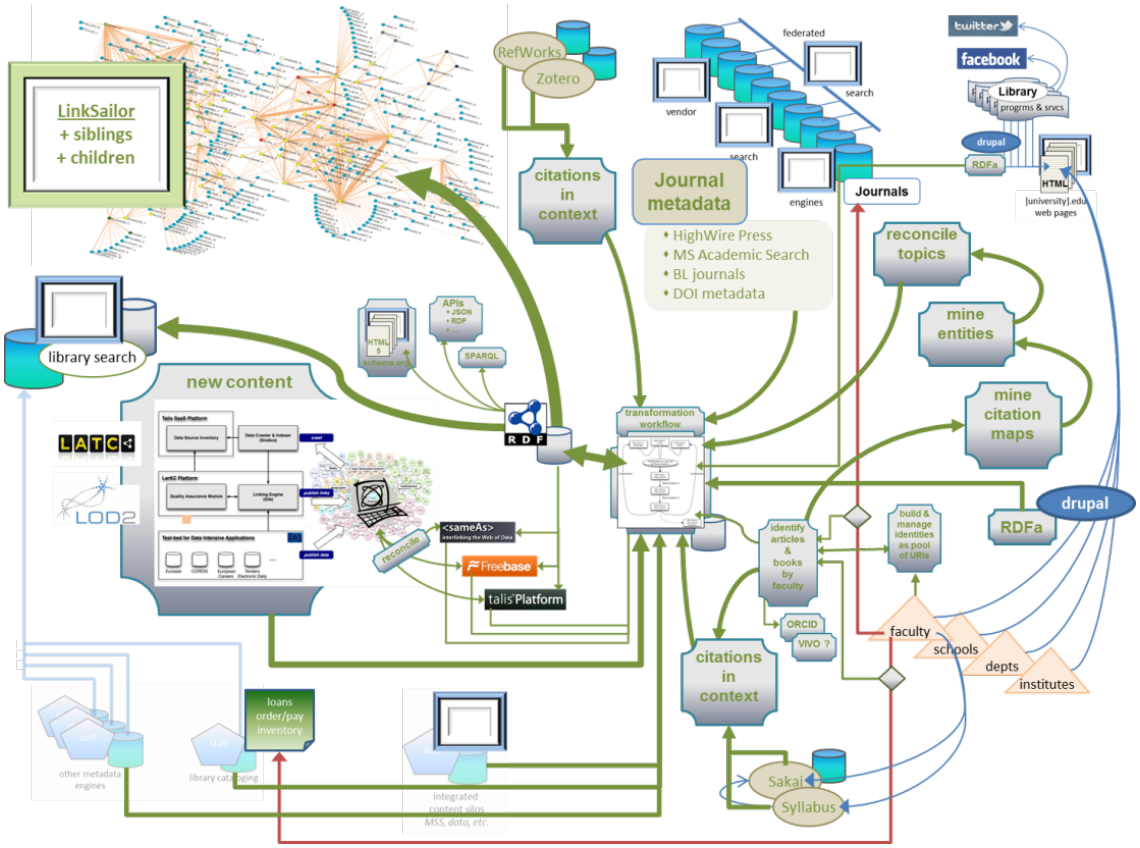
- RDF data store(s) and systems
  - here we plan to draw on the extensive work underway in LOD2 (Linked Open Data 2) projects, work that began in 2010 and is funded at 6+M€ through 2014. Many of the capabilities that this project will require are available or under development in the LOD2 technology 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
  - Open University
    - especially the LUCERO project
  - 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
  - pursue "authority record" for every name (as DNB does for Germany)
  - British Library authorities included with US data
  - pursue access to DNB and BnF data
- VIAF
  - 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

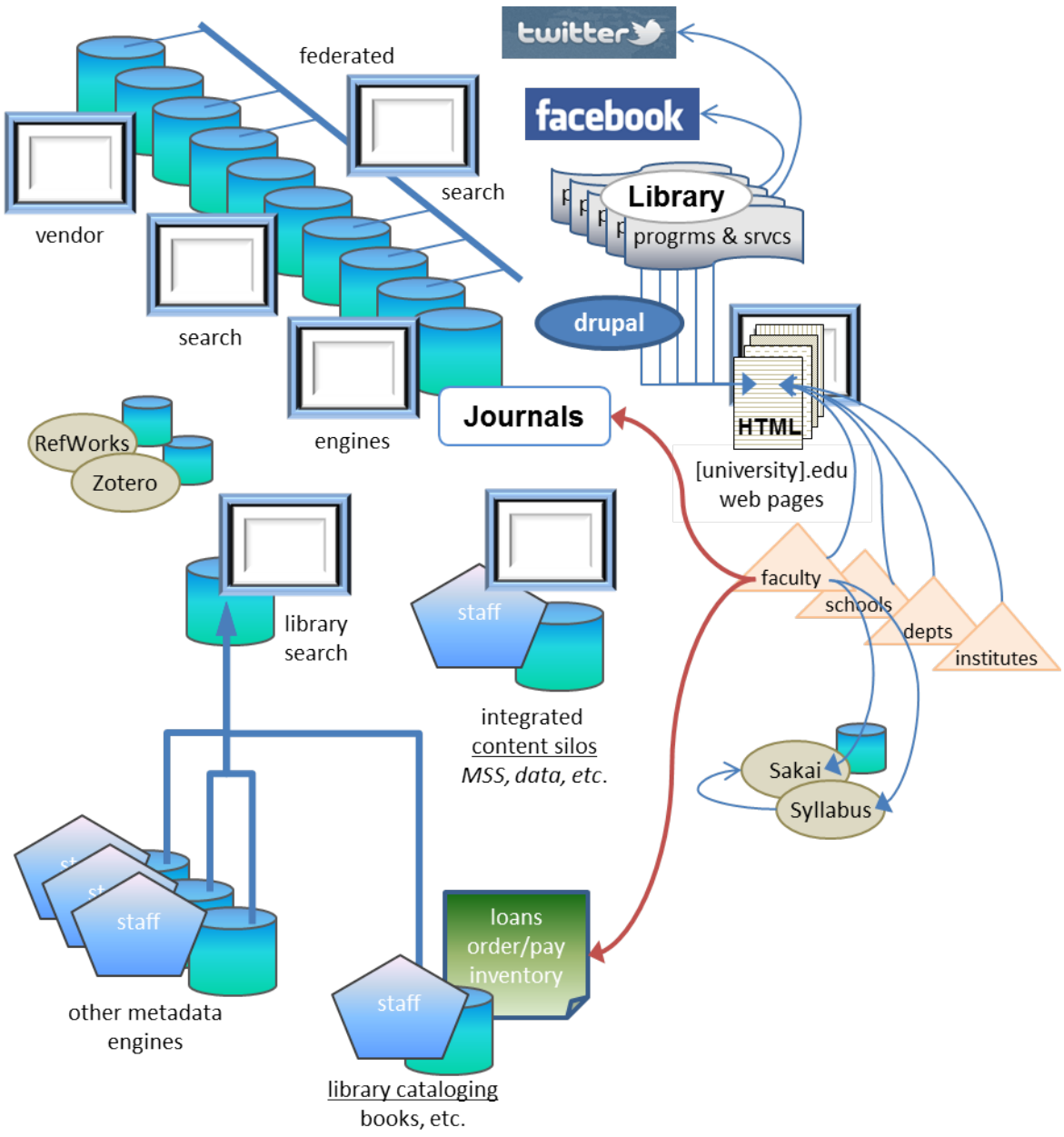
- Partner with the eXtensible Catalog project building and elaborating utilities to create an open source engine for transforming MARC data to RDF triples conforming to the project's data model

Potential Partnerships

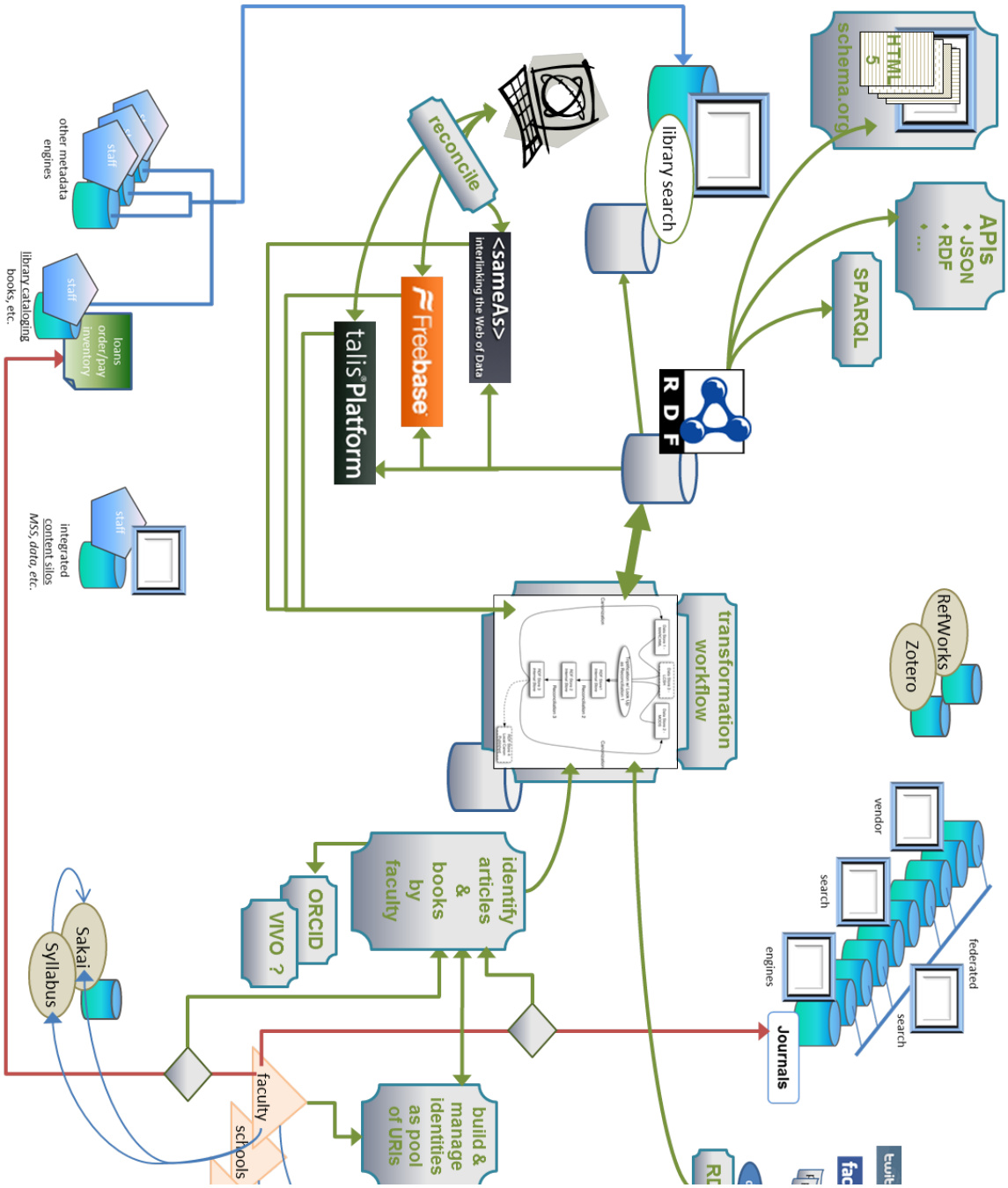
1. Talis
    - Talis Consulting
        - strategic planning
        - development of data model (they consulted with BL)
    - Talis Platform and LATC Interlinking Platform
        - exposure of project's linked data products in web-wide venue
        - reconciliation of URIs to enhance local pool of 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

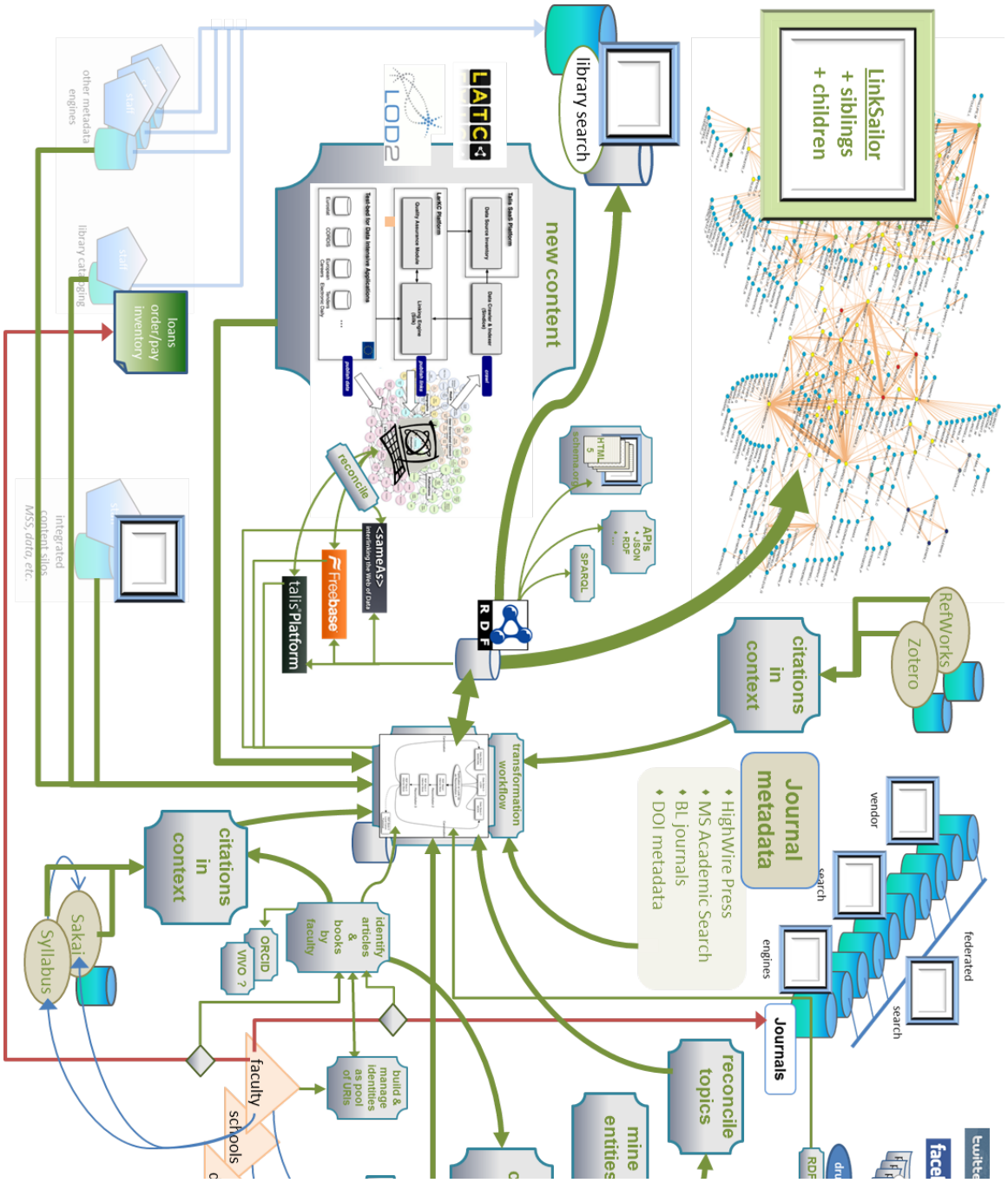
1. Schematic snapshot – current characteristics



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.

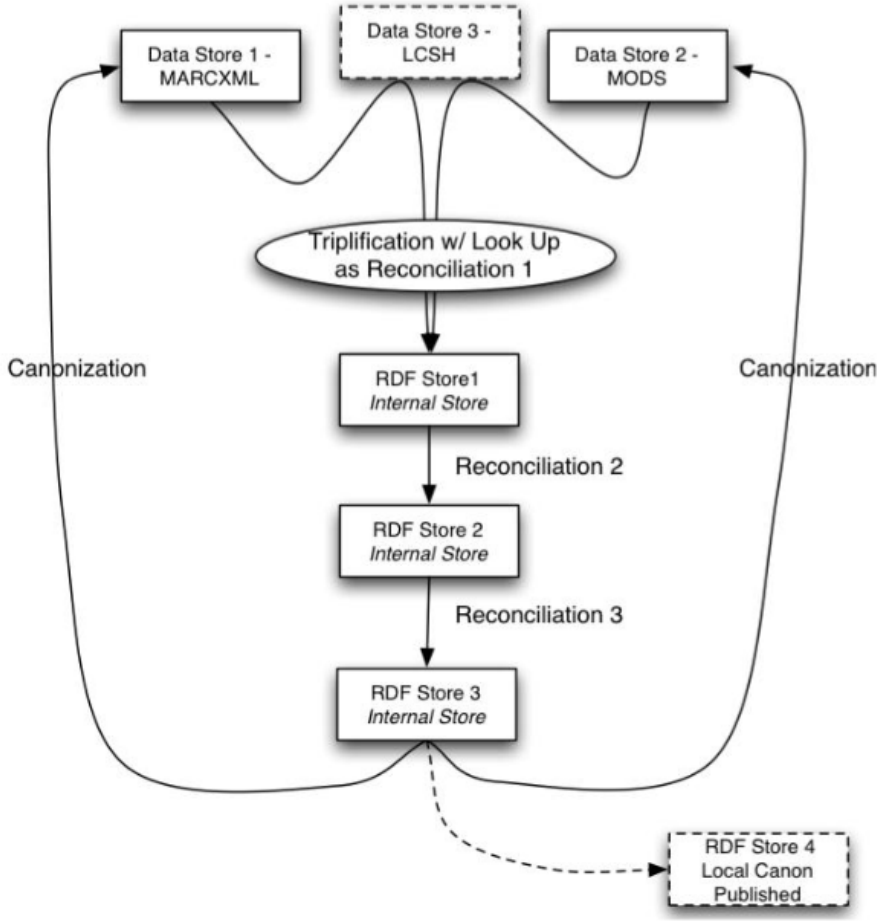
3. Leave linking to later

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:



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.

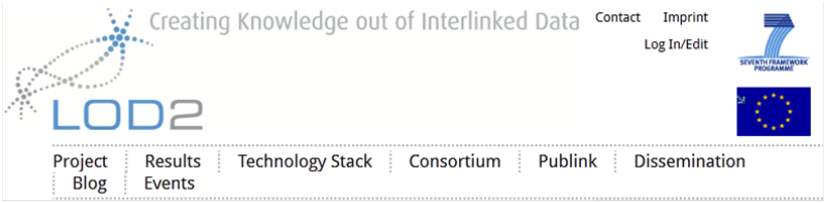
2. The whole process will be replayed on a continuous basis, as more data arrives in the Data Stores. It is likely that the simplest way to do this is to do the recapture (with canonization). Since the reconciliation information is out with the stores, it will still apply to the newly recaptured RDF.

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.

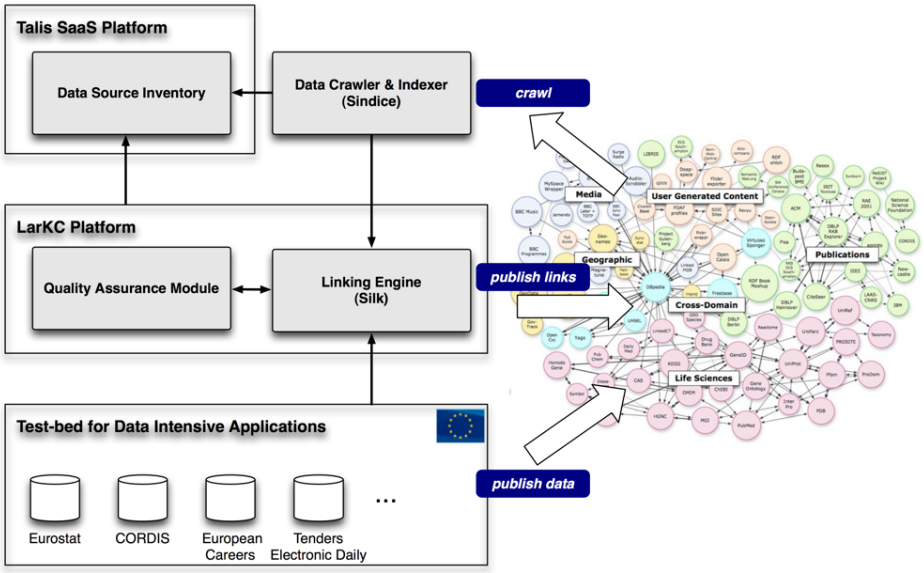
WIQA
The Web Information Quality Assessment Framework is a set of software components that empowers information consumers to employ a wide range of different information quality assessment policies to filter information from the Web. Information providers on the Web have different levels of knowledge, 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

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: