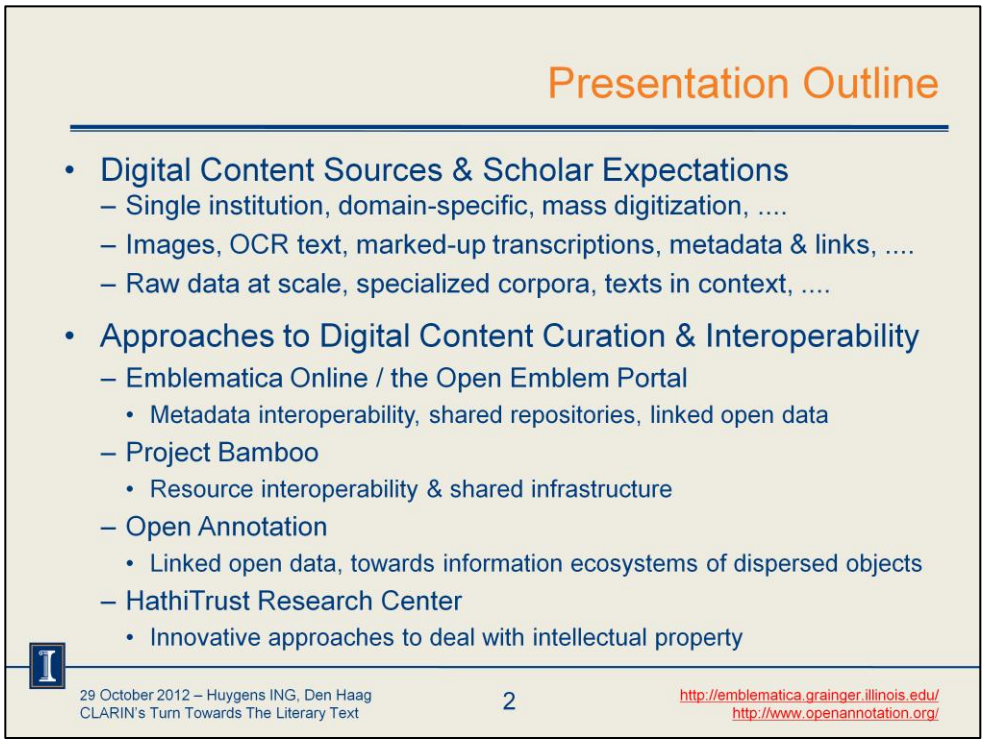
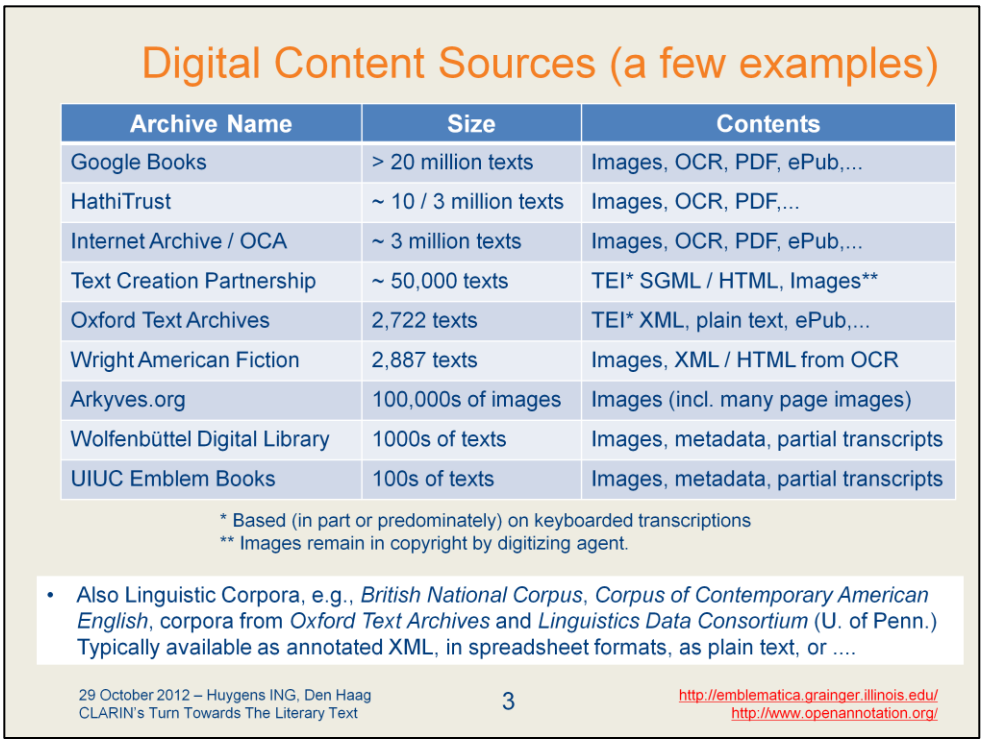
My colleague Harriett Green and I are librarians at the University of Illinois at Urbana-Champaign and lecture from time to time in the Graduate School of Library and Information Science at Illinois. Accordingly, we approach the issue of providing more integrated access to digital resources from the perspective of librarians and information scientists as compared to how a scholar working in literature, linguistics or history might approach the problem. Still we engage with such scholars on a regular basis, and it is our objective to breakdown the boundaries between digital content silos and repositories in a manner that will help scholars achieve their research and pedagogical goals and ambitions. In recent years we have individually and collectively worked on several projects that have content interoperability and integration as a goal, and my comments today draw from our experiences on these projects, i.e.: the Emblematica Online project through which I met Els and which has been carried out in collaboration with HAB and support from both the NEH and DFG; Project Bamboo, a multi-institution digital humanities infrastructure project supported by a generous grant from the Andrew W. Mellon Foundation; the Open Annotation project, another multi-institution project supported by the Mellon foundation; and the HathiTrust Research Center, a collaboration with Indiana University and the University of Michigan.

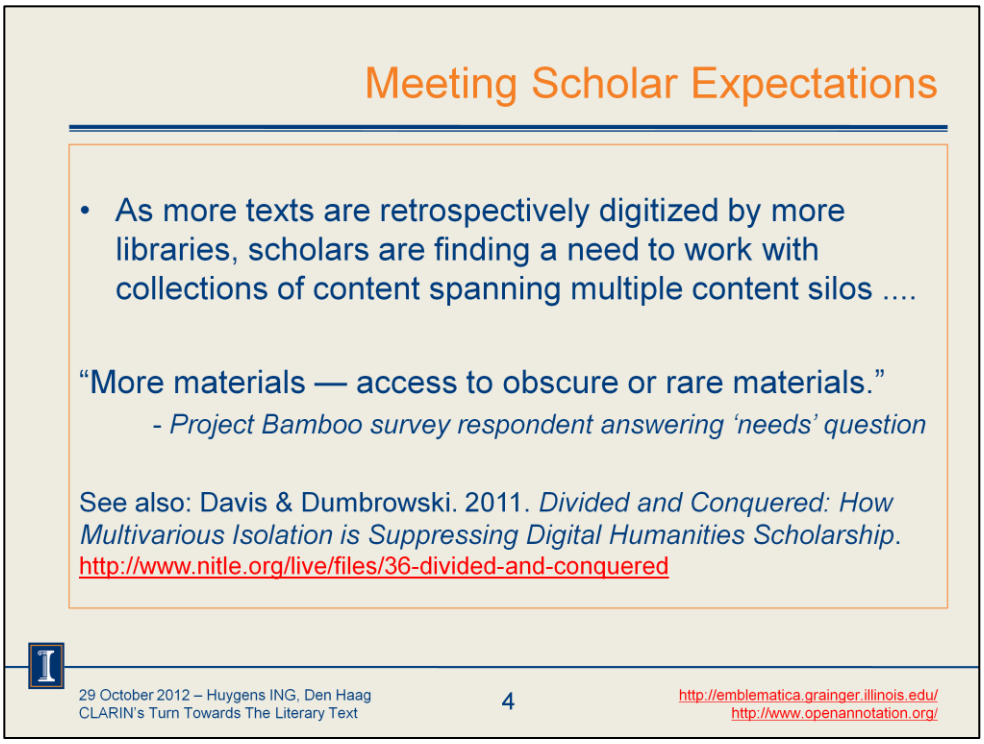


I plan to start today by discussing briefly the breadth of scholar requirements in regard to integration and interoperability of retrospectively digitized content. This is a very broad topic, and I one that I do not have time to address comprehensively, but in order to help provide context and motivation for the comments that follow, I do think it useful to dip quickly into a few illustrations of scholarly requirements and talk generally about a few of the use cases against which different visions of digital content silo integration must be measured.

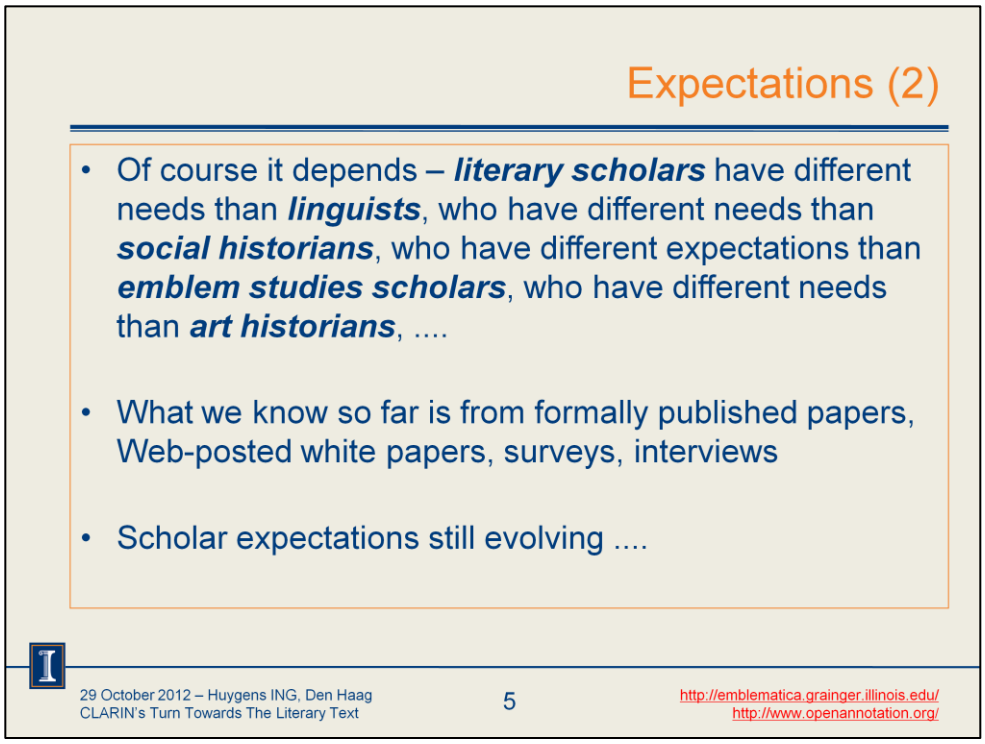
From there I will illustrate some of the different approaches that are being developed and tested, drawing illustrations from each of the projects I mentioned above. There are not surprisingly a range of different approaches and metaphors in play right now, of varying degrees of complexity and sophistication – starting with interoperability at the metadata level, moving up to interoperability and integration of content and content models within a larger centrally-defined infrastructure, to more dynamic models of content resources that essentially create distributed objects and threads of scholarly discourse that span repository and silo boundaries, and finally to some innovative approaches that bring tools to resources enabling non-consumptive analysis in a way that respects intellectual property restrictions.



It is important to acknowledge at the outset that repositories vary greatly in size, content and format availability, whether full text is created relying on machine OCR or by humans, metadata completeness, amenability to machined-mediated interactivity, etc. Most do not have robust, explicit APIs (HathiTrust is a notable exception), although many have de facto RESTful services that can be used like an API. Most use idiosyncratic identifier schemes and included little if any linked open data in metadata records. Some (e.g., linguistic corpora) are meant to be downloaded (i.e., as a snapshot) and used locally.



Which brings us to the other side of the coin – having all this content how useful in availability and presentation in meeting scholar expectations. Humanities scholars are increasingly aware of and interested in using digital resources in teaching and pedagogy. However, the heterogeneity of online text and text/image data sources remains an obstacle. Scholars want to be able to discover resources without regard to silo repository boundaries, and they want to be able to integrate into research and teaching corpora texts from multiple different silos. The isolation of a scholar with his or her handful of texts is no longer the only model of humanities scholarship. Increasingly humanists are finding benefit in sharing tools and content and even working collaboratively at large scale projects involving digitized resources.



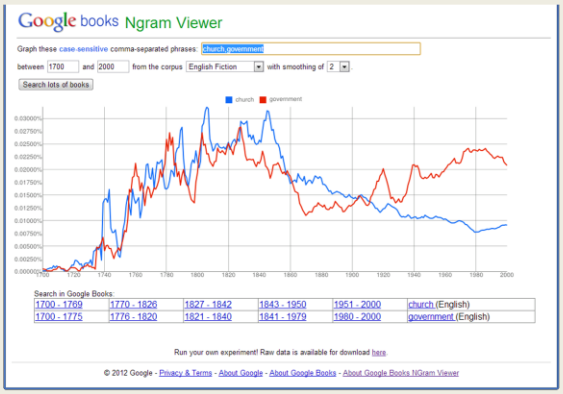
Of course humanities scholars are not monolithic in their research requirements and expectations. This has implications for how we coordinate discovery of, access to, and use of retrospectively digitized texts. The last few years has seen the publication and Web posting of several very good studies and white papers summarizing a wide range of use cases and the needs and expectations of scholars working in a variety of domains. Many of these papers are based on small, localized samples and anecdotal evidence, but we are beginning to see raw data from an number of multi-institutional studies, be they still modest in sample size. And scholar expectations are still evolving as scholars learn more about what they can do with digital resources.

Expectations (3)

Some scholars want large volumes of raw text, noise and all.

Some analytical tasks require content at scale.

Michel, et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books [Google Ngram Project]. *Science* 331:176.



The next 3 slides illustrate the breadth of use cases and scholarly domains we need to support. Here using the example of the Google Ngram Project is illustrated one class of use case -- the linguist interesting in applying analytic tools to raw full-texts from literally millions of texts. In this context scale is particularly important. Some noise in the data (e.g., as is common with OCR) is tolerable. But even here, being able to break down the data accurately as regards genre (fiction in this example) and publication date – both implying at least minimal metadata – is important.

Expectations (4)

Some scholars need richly described & linked primary sources

Audenaert & Furuta. 2010. ***What Humanists Want: How Scholars Use Source Materials***. Proceedings of the Joint Conference on Digital Libraries

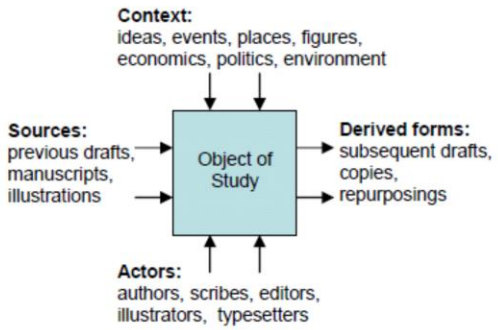


Figure 1: SCAD: Sources, Context, Actors and Derived Forms

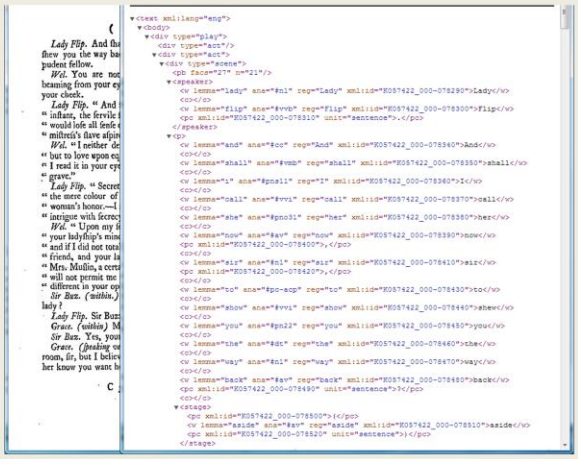
Others, e.g., social historians and literary scholars studying the relationship between literature and contemporary culture, are satisfied with considerably smaller sample sizes, but require very rich metadata, including relationship metadata and metadata not always found in standard bibliographic descriptions, e.g., the gender, nationality and religion of an author.

Expectations (5)

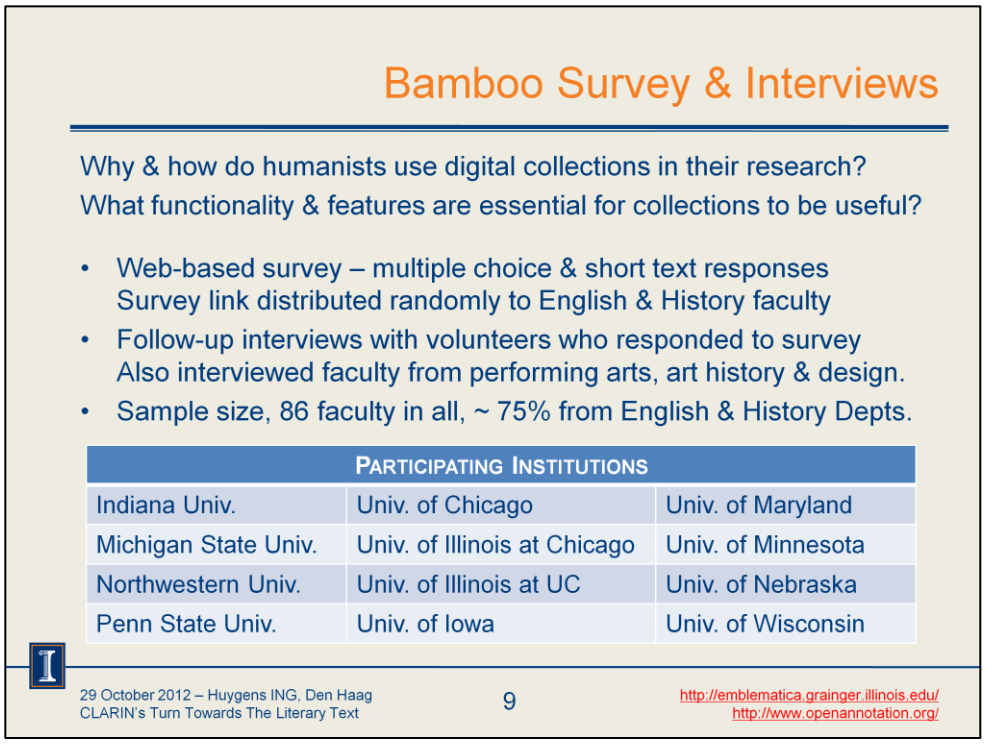
Some scholars need carefully formatted & enriched texts

Rowe & Mueller. 2012. ***Estimating F21 curation rates on a "Rough Cut / Fine Cut / Rich Curation" model:***

"We begin with a TEI P5 processed corpus of ~800 plays from the Text Curation Partnership (TCP) and focus on completing incompletely transcribed words ...."

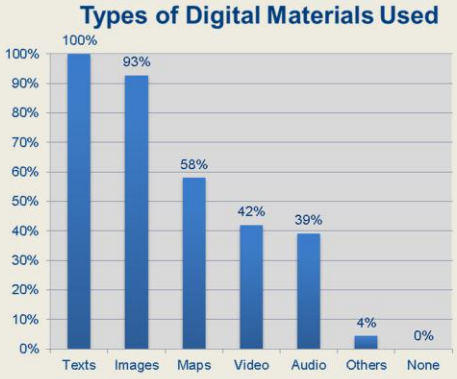


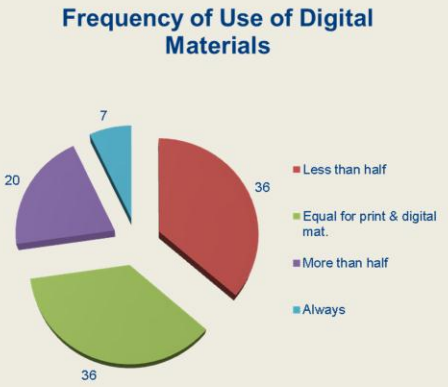
For many scholars as well the noise of OCR is not acceptable, and it is important to delineate in text recognized intellectual structure, such as for a digitized play script the importance of digital copy provenance and distinguishing between dialog, the attribution of dialog, and interspersed stage direction. For such use cases the process of going from page image to marked-up text adequate for certain kinds of analysis, while largely machine mediated, often mandates some human intervention. When thinking about repository interoperability, the question becomes where in the workflow is this human intervention inserted, and if after materials have initially been deposited in a repository, how and where are the outputs of human-mediated text enhancements saved and how are they provenanced?



The importance of these considerations is becoming more and more apparent as we continue to survey and talk to scholars. These next 5 slides provide a sampling of results from some survey and interview work done in part in connection with my institution's involvement in Phase I of the Bamboo Technology Project. Our goal was in particular to inform the Project's Phase 2 approach to collection development and prioritization (i.e., for integration into the Bamboo infrastructure), but we learned as well a lot about how humanities scholars are using and want to use digitized texts and images.

Survey Respondents Use of Digital Collections





Green, Harriett. "What Are You Going to Do with That Data?: A Needs Assessment of Scholars and Digital Collections." Paper presented at the *KU Digital Humanities Forum 2012*, Lawrence, KS, September 22, 2012

Survey Results – General Requirements

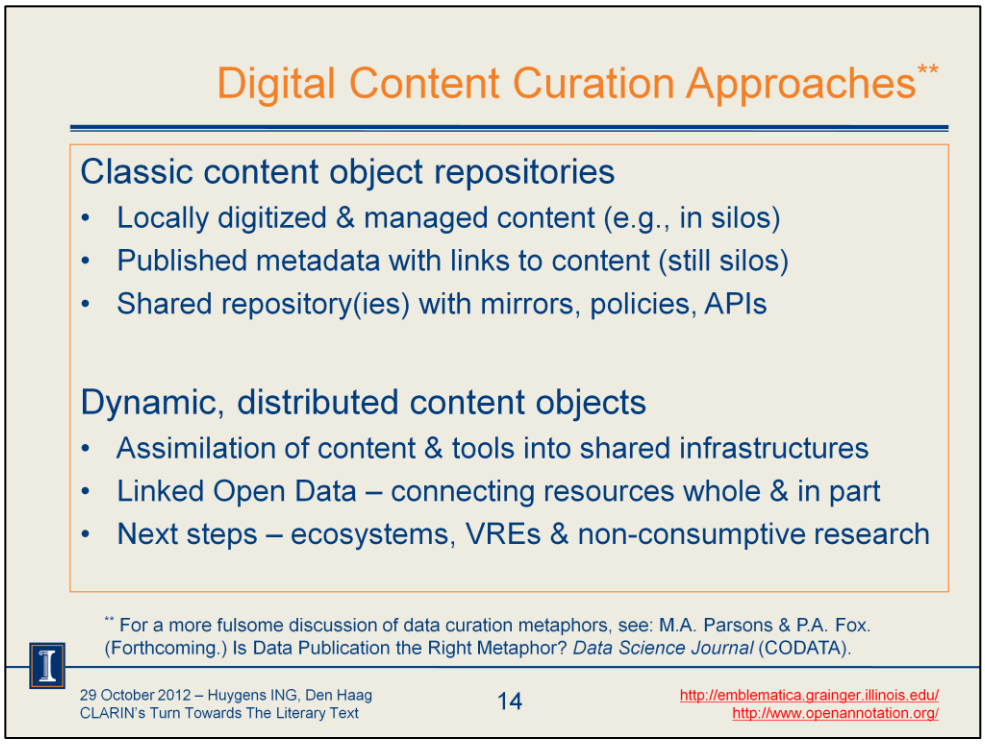
- “Ability to create one's own "library" held on the digital database's server space (like Hathi Trust collections allow) -- a concern for those of us who don't want to eat up all our hard drive space with PDFs.”
- “exporting files and creating my own text and visual files either for teaching or research purpose[s]”
- “high resolution (I do a lot of on-line work and my eyes are really degenerating)”
- “More materials—access to obscure or rare materials.”
- “Ability to know copyright issues or costs up front.”

Survey Results – Requirements for Text

- "links to related material or bibliographic resources"
- "metadata must include rigorous bibliographic information and stated principles, either explicitly derived from existing modern critical editions or stated as unique to this database."
- "ease of access to items within collection (not half a dozen mouse clicks to get to it)"
- "capability to be read on multiple devices (e.g., iThing, laptop, desktop)."
- "The ability to annotate these texts in private and public environments, where I could keep stuff for myself or share it with others."

Survey Results – Requirements for Images

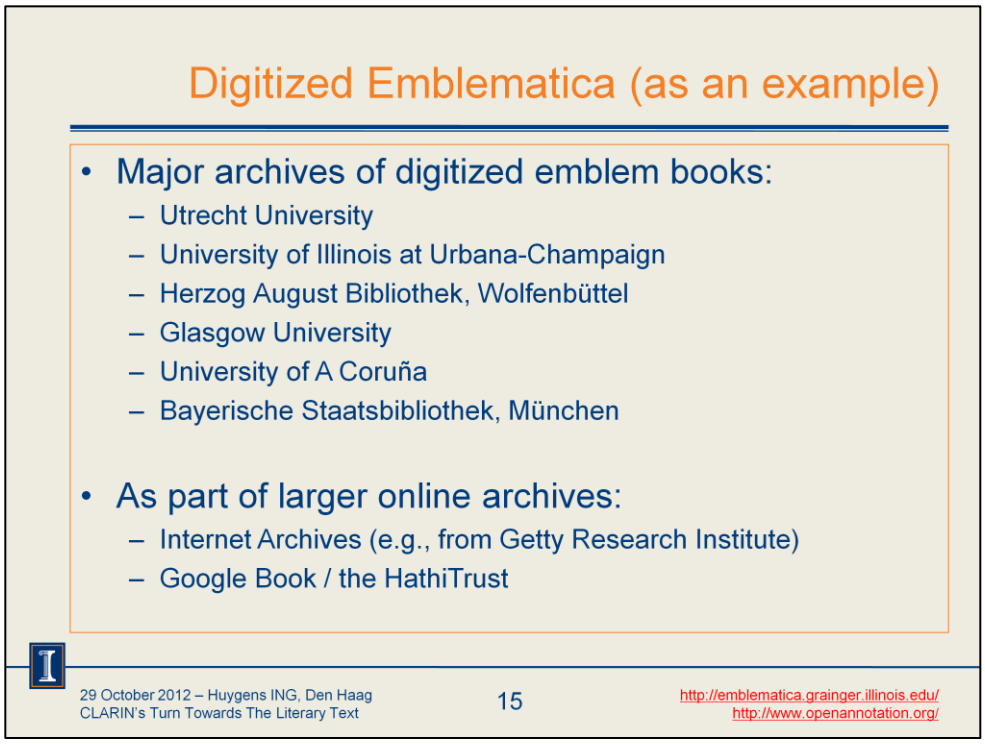
- "Ability to export in multiple formats (i.e. jpegs and higher resolution tiffs)"
- "information on permissions and how to request them"
- "ability to have a folder of images in the collection to which I can return that contains all of the copyright and original publication material associated with the image"
- "viewing and zooming tools, color accuracy, ability to export, reliable metadata"



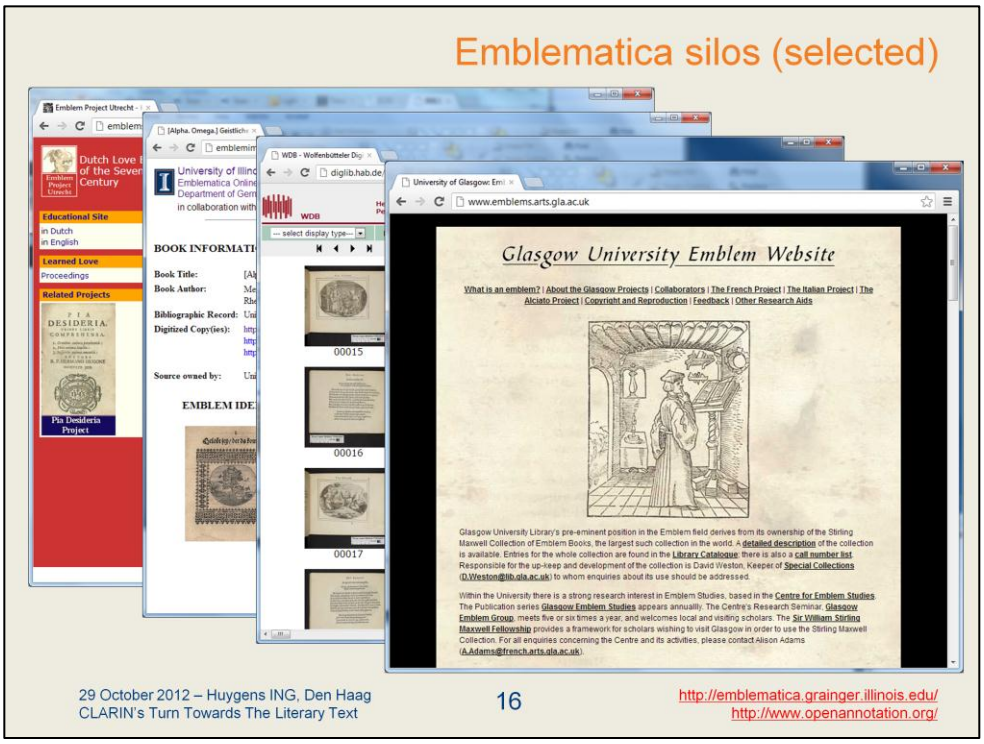
Just as scholar expectations have evolved over the last decade with respect to the use and utility of digital resources, so have library and curator roles evolved with respect to digital resource dynamism, services and interoperability. Understandably the initial metaphor for digital libraries was the book on the shelf. Libraries know how to acquire, process and maintain print books on library shelves. So initially the focus of retrospective digitization was the creation of the printed book analog on a digital shelf – *digitization as preservation* in the terminology suggested by Peter Daly. This metaphor, however, does not take advantage of the capabilities offered by the digital format and the infrastructure of the World Wide Web. Daly suggested *digitization as enrichment*, which though imprecise is as good as most any other terminology suggested. Digitization as enrichment assumes more complex workflows, tools, and services that make retrospectively digitized content more accessible, more discover, and ultimately more useful than the content was before digitization. In the case of digitized emblematica this has meant more granular access and description (i.e., at the level of individual emblems and emblem components) and the potential at least for better linkages between resources and a recognition that metadata in particular and resources in some cases must be seen as dynamic. This in turn has implications for how we approach content organization and curation. The initial goal of simply creating online analogs to a local collection gave way quickly (more than a decade ago) to a recognition that resources needed to "published" on the Web. Initially this meant making metadata available to be 'pulled' by interested parties. The content itself remained in local collection silos, sometimes isolated even within a given institution. Gradually this has given way to models that assume centralized, shared repositories to hold content. These may be domain or consortially set up, managed and paid for. As a preservation strategy these can be cost-effective when mirrors and preservation policies are clear. These also can be a good first step towards more robust infrastructure if bolstered with a good and well-thought out set of HTTP-based application programming interfaces, such as used

by the HathiTrust. Even better is when mirror technologies, policies, and APIs reflect community consensus and make use of community-accepted standards.

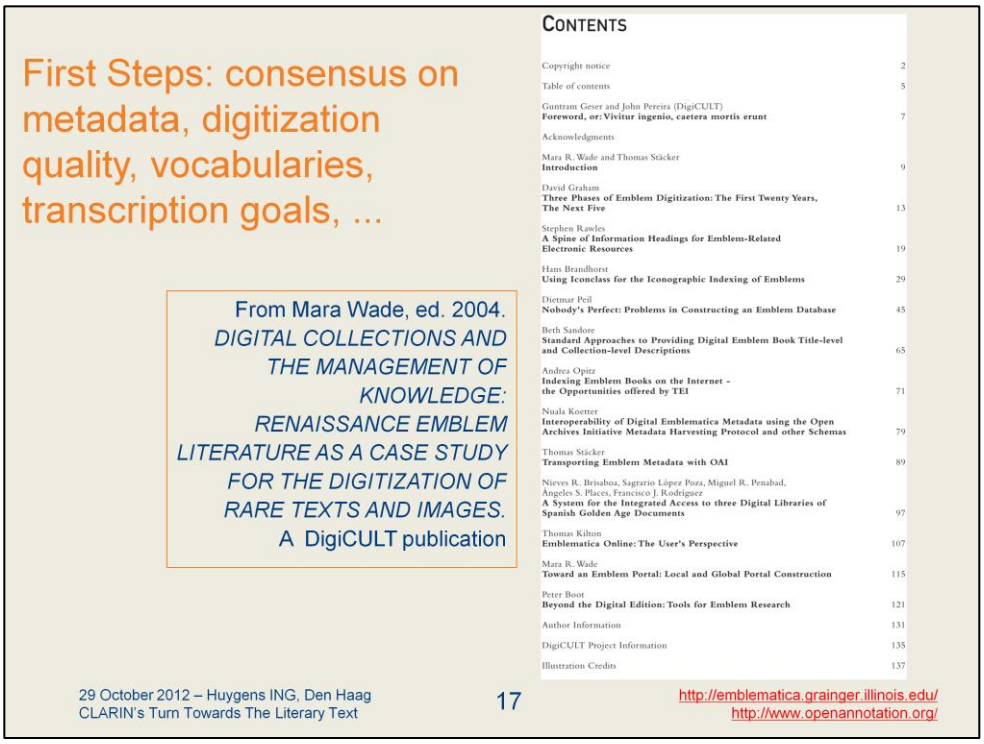
However, even as we have learned that the book-on-the-shelf metaphor is inadequate in the context of today's Web, so are we learning that retrospectively digitized texts and their associated metadata are not as static and unchanging as we once thought. As illustrated above, text and images come in a multiplicity of formats. Texts can be created by OCR or by keyboarding, but once created they can be further enhanced and enriched by being marked-up or being annotated. A digitized book can be referenced, used, and re-used as a whole or in part. Mashups can bring together parts of a digitized resource into a new virtual resource residing in a wholly different repository. We now need to deal with multi-part, dynamic, potentially distributed, potentially recombinant content objects. This has required new and innovative approaches to curation of, management of, and access to digital content.



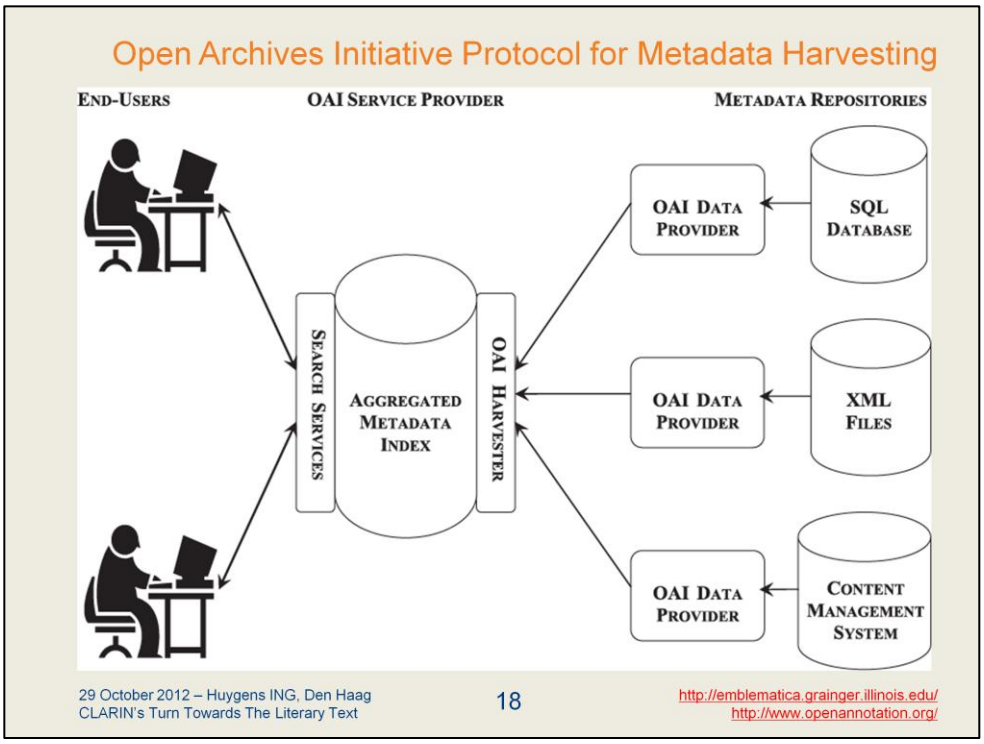
Consider as an example the history of retrospectively digitized emblematica. The initial focus was on the local digitization and mounting of high-valued individual collections held by individual institutions. Over time the community has developed and implemented consensus for digitization and metadata standards. We are now poised for reciprocal sharing of metadata and even content, albeit for the limited purpose initially of dark archiving. To this point the metaphor of publication and digital library shelves predominant, but we are also well-positioned because of this foundation to begin experimentation with more advanced sharing, and in particular with linked open data.



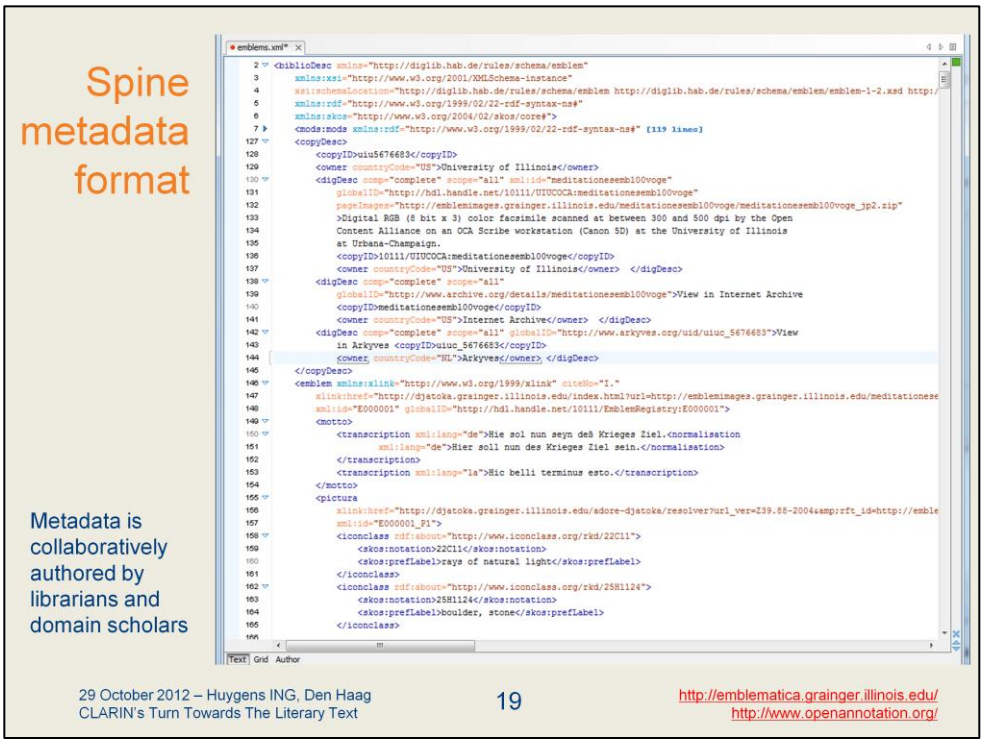
Here are screen shots from 4 of the major digitized emblematica collection portals. These sites are very different one from another. However, because of nearly a decade of close communication facilitated in part by the Society for Emblem Studies, behind the scenes these sites are much more similar than dissimilar and are well-positioned to interoperate.



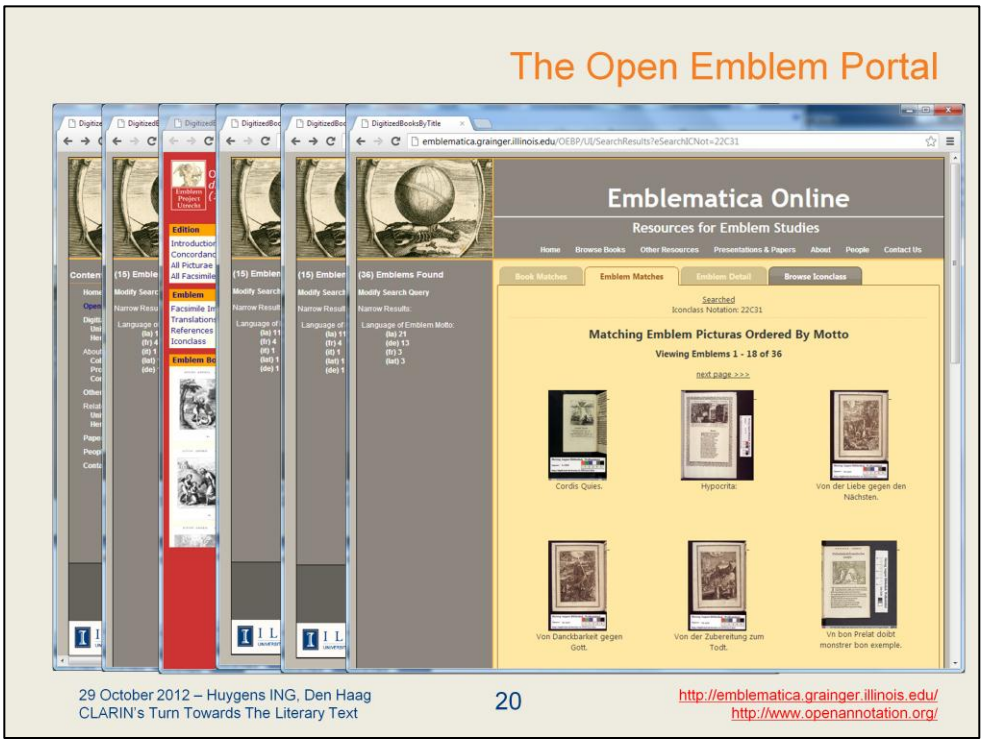
Key has been the first steps – consensus on foundational features and capabilities. This volume that came out in 2004 is emblematic (pardon the word choice) of the community-based approach taken to creation and curation of digitized emblem resources. Some very important ideas captured here – David Graham's overview of emblem digitization, Stephen Rawles's Spine metadata proposal, Hans Brandhorst's paper on the pertinence of the Iconclass vocabulary for emblems, Thomas Stäcker's paper on the potential of OAI-PMH for this community, Thomas Kilton's paper on the importance of addressing user requirements, Mara Wade's vision of a global open emblem portal (which is finally near realization), Peter Boot's paper looking forward at tool's for emblem research. Insights such as represented by these papers allowed the emblem community to progress from isolated silos towards more interactivity and interoperability across silos.



One of the technologies that's been used in this process is the Open Archives Initiative Protocol for Metadata Harvesting. As discussed later, this will likely be superseded in coming years by ResourceSync and other de facto standards such as AtomPub and the related CMIS standard, but nonetheless, an early focus on interoperability and OAI-PMH helped to drive consensus on metadata format. At a minimum OAI-PMH greatly facilitates awareness and discovery spanning multiple silo repositories.

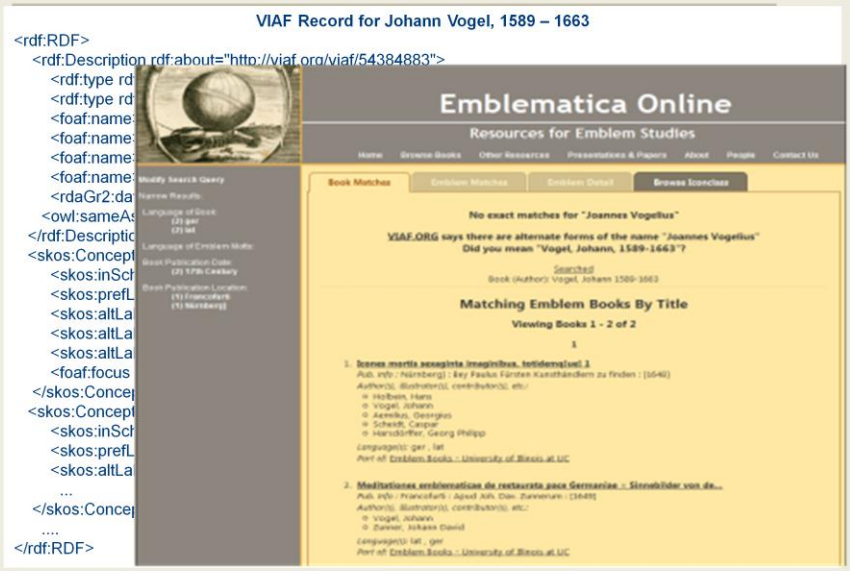


But OAI-PMH also illustrates that technology alone is insufficient. Emblem literature highlights the opportunity digitization affords to go beyond the book metaphor. Just as the 1967 publication of the encyclopedic *Emblemata* volume by Henkel and Schöne invigorated emblem studies by providing for the first time a resource that index a corpus of emblem books at a level more fine-grained than the containing volume. *Emblemata* provided access to individual emblems, an index of emblem mottos, and an index of meanings. Users could browse this massive print resource according to subject headings, such as the "heavens," "mammals," "trees," and so on. In the wake of this path-breaking publication Emblem Studies grew at a heretofore unprecedented rate. So in digitizing emblem books it was natural to consider the possibility of making digitized emblem literature discoverable not only at the level each volumes bibliographic description, but also at the level of each emblem, emblem motto, and emblem pictura descriptors. This recognition led directly to the creation of the Spine metadata format, based on the publication of the paper entitled a Spine of Information Headings for Emblem-Related Electronic Resources. Key to the successful implementation of this new metadata format were the decision to leverage as much as possible existing metadata standards and the recognition that Spine metadata records would need to be created in some measure through collaboration between librarian-cataloger and domain scholar.

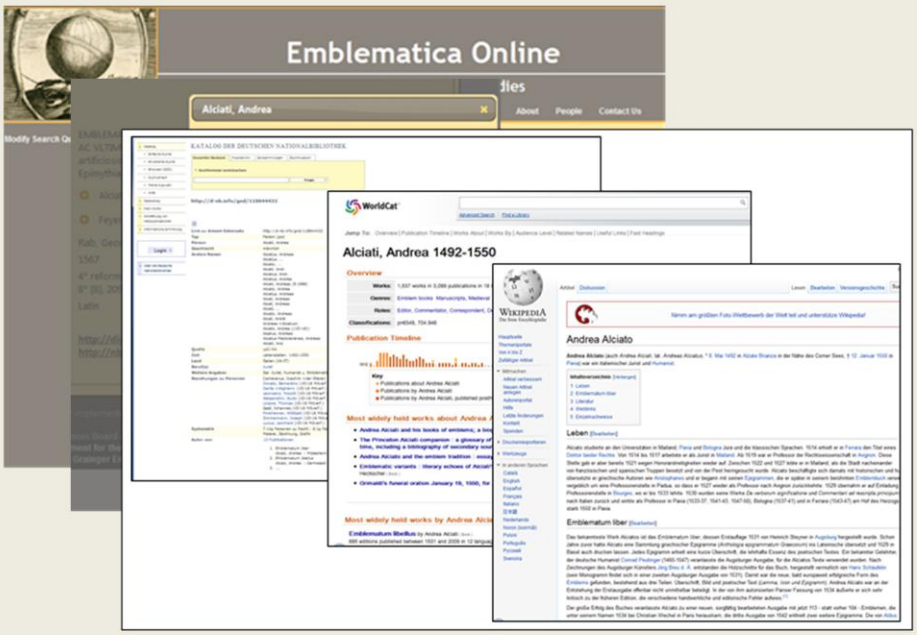


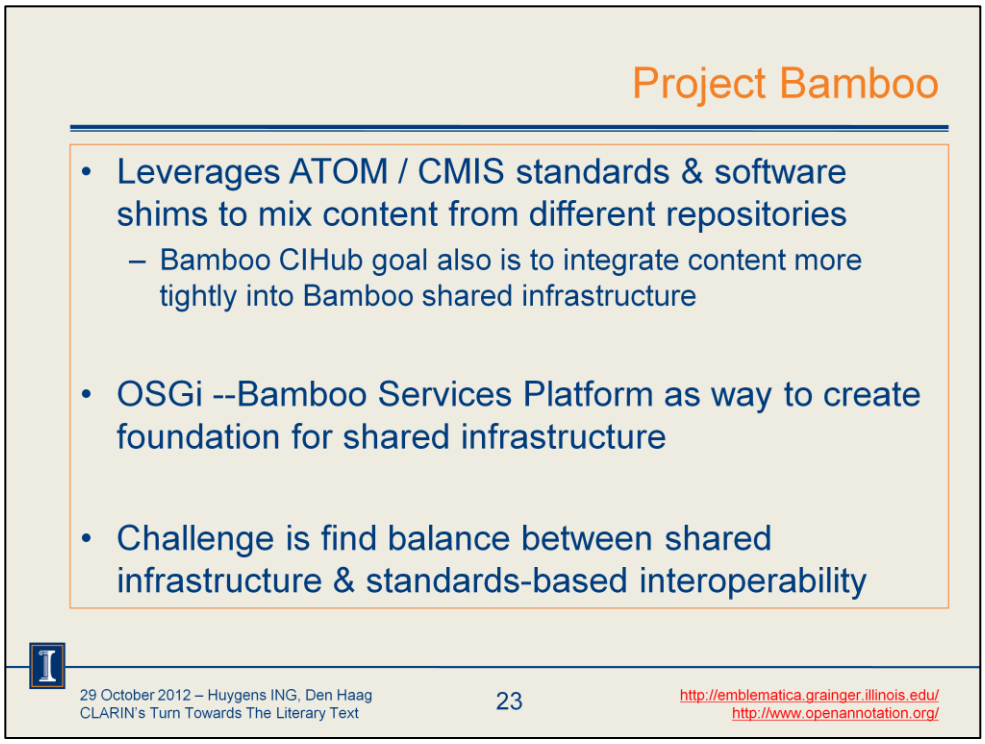
The Spine metadata approach then enables not only interoperability, but interoperability at a particular useful level, reflecting the granularity requirements of the core user community.

Possible Future Directions for *Emblematic Online*



E

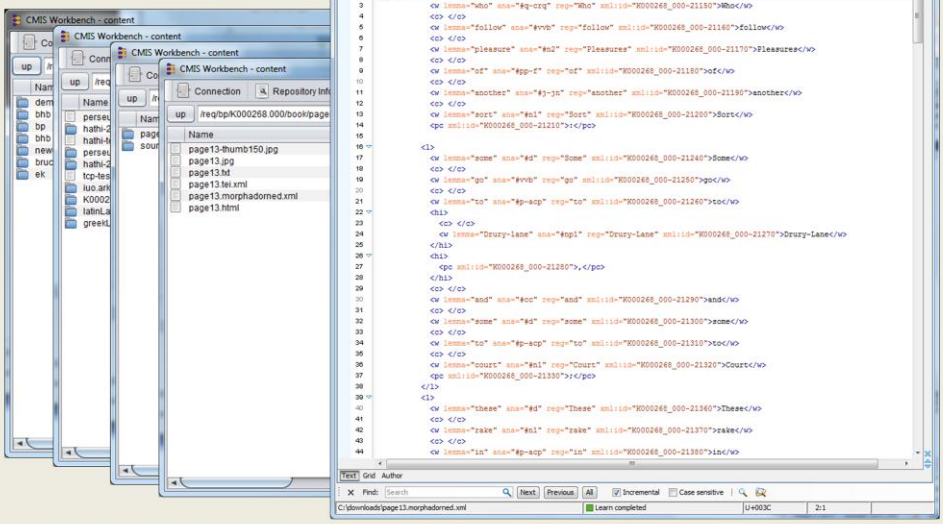

Emblematica Online
Alciati, Andrea
About
People
Contact Us
KATALOG DER DEUTSCHEN NATIONALBIBLIOTHEK
WorldCat
Alciati, Andrea 1492-1550
Overview
Publication Timeline
WIKIPEDIA
Andrea Alciato
Leben
Emblematum liber



Bamboo CIHubs implement Apache Chemistry

Bamboo BSP implements Apache ServiceMix

CMIS* View of Content



*Content Management Interoperability Services https://www.oasis-open.org/committees/cmis/

Other BSP services

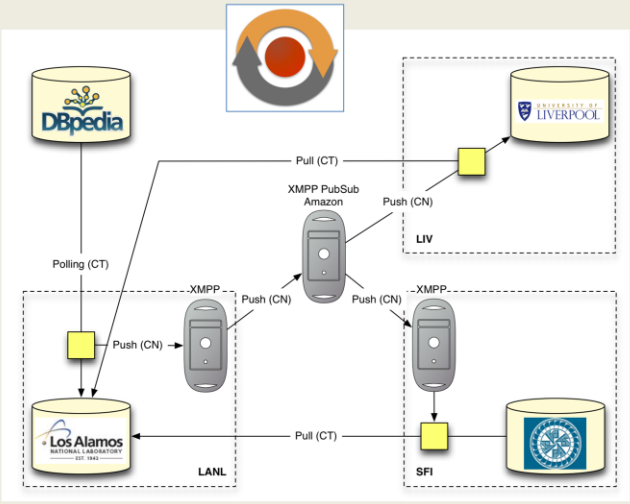
- Identity Management (IdM)
  - Authentication Service: confirm Bamboo Person ID
  - Bamboo Person ID allows access to content, tools, ...
- Content remediation and remodeling
  - One tool requires plain text by paragraph
  - Different tool requires TEI with POS tagging
- Transient stores for annotations & other derivatives
- Drupal-based Research Environment
  - Prior experimentation with HubZero, Alfresco, VRE

A ResourceSync* Precursor Experiment: Testing a Push/Pull ResourceSync Approach

Two Destinations (right) synchronize with a LiveDBpedia instance (bottom left) by acting upon change notifications to which they subscribe.

From. *ResourceSync: Towards A Web-Based Approach For Resource Synchronization*.

A Project Briefing given Herbert Van de Sompel at CNI 2012 Spring Membership Meeting.



* For more information see: Herbert Van de Sompel, et al. 2012. A Perspective on Resource Synchronization. *D-Lib Magazine* 18 (9/10). doi:10.1045/september2012-vandesompel

26

From D-Lib Article: At the core of the resource synchronization problem is the need for one or more Destination servers to remain synchronized with (some of the) resources made available by a Source server. Three distinct needs are recognized:

- **baseline synchronization**: An approach to allow a Destination that wants to start synchronizing with a Source to perform an initial synchronization operation.
- **incremental synchronization**: An approach to allow a Destination to remain up-to-date regarding changes at the Source.
- **audit**: An approach to allow a Destination to determine if it is in sync with a Source.

From CNI Briefing: In order to explore the ResourceSync problem, a straw man resource synchronization approach was formulated and tested earlier this year. The test is illustrated here:

- A Source **pushing** change notifications to Destinations using XMPP PubSub as a carrier protocol – Push CN in the Figure.
- A tweet-like **change notification language**, deliberately chosen to be writeable and readable by both machine and human agents –
  - ***http://megalodon.lanl.gov/dbpedia/data/Paris updated at="2012-03-05T19:54:39Z" #dbpedia $resync***
- A Destination **pulling** the entire resource about which it received a change notification from the Source – Pull CT in the Figure.

See presentation and article for further details and an in-depth discussion of the technical issues.

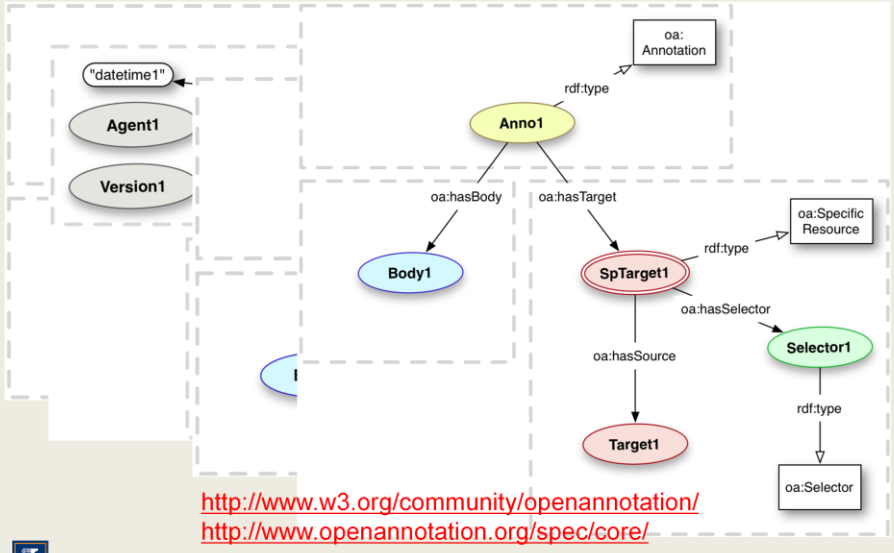
Support for Annotation, Scholarly Discourse & Netchaining*

Web resources are dynamic & have dynamic relationships
New derivatives & related resources are created over time
***This is key to the vision of a* Semantic Web**

- To enable annotations to become part of the Semantic Web:
  - Need a Web and Resource-centric, RDF-based, data model & ontology
  - This is the goal of the W3C Open Annotation Community Group
- By giving distinct identities to annotations & annotation bodies, enable *NetChaining* & other forms of annotation-based discourse

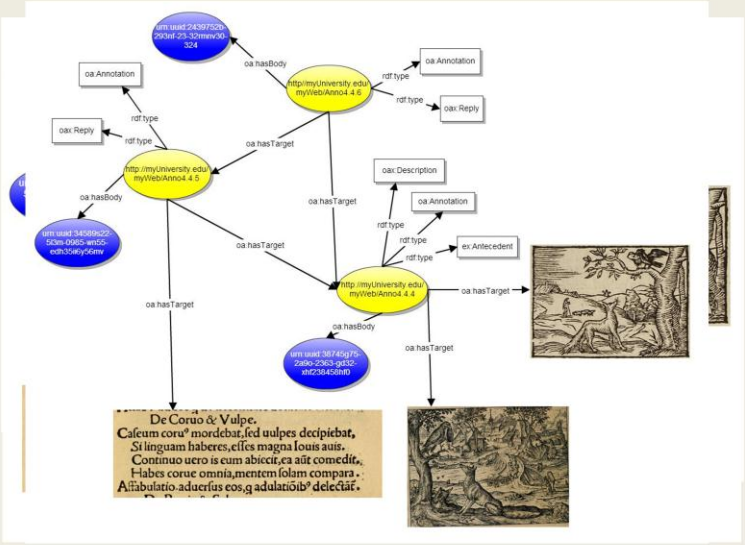
* Suzana Sukovic. 2011. E-Texts in Research Projects in the Humanities. *Advances in Librarianship* 33:131-202. Suzana Sukovic. 2008. Convergent Flows: Humanities Scholars and Their Interactions with Electronic Texts. The Library Quarterly 78 (3): 263-284.

The Open Annotation Data Model



http://www.w3.org/community/openannotation/
http://www.openannotation.org/spec/core/

Annotation-based discourse illustration

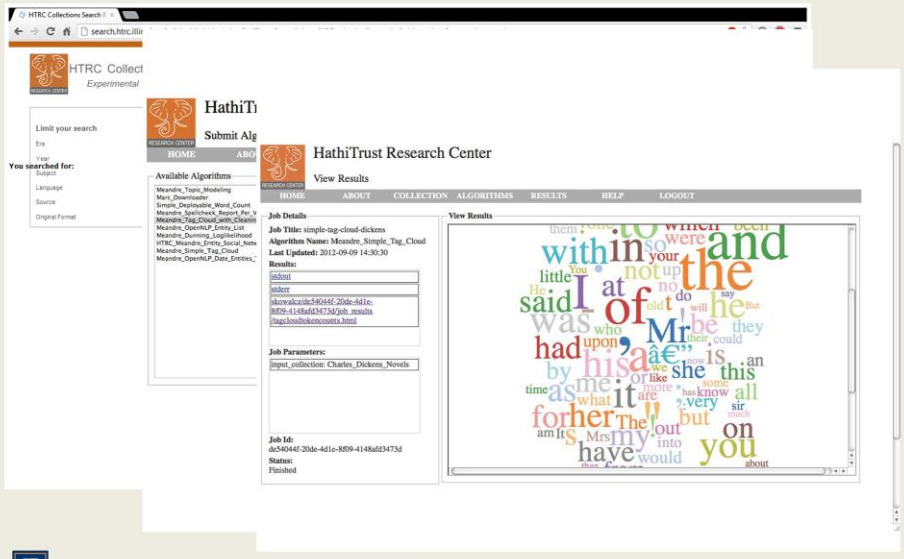


HathiTrust Research Center

Research Arm of the HathiTrust

- Enables computational access to HathiTrust Texts
  - Bring the tool to the content when restrictions on content access precludes taking content to the tool
  - Supports ***non-consumptive*** research & analysis
  - SEASR / Meandre built-in
- Tool builders can write to published HTRC API
- Opportunity for large-scale computation

HTRC Non-consumptive research: bring tools to content



Concluding Thoughts

- Content Providers:
  - Need to provide open metadata & persistent object URIs
  - Must recognize the scholarly case for re-use & new derivatives
- Service Providers:
  - Need to maintain provenance & reproducability
  - Must avoid duplication when unnecessary
- Open research questions (a sampling):
  - How can we de-dup & FRBRize overlapping collections better?
  - How do metadata, annotations, corrections propagate upstream?
  - When should we link vs. cache vs. harvest vs. ResourceSync?
  - What are the limits of non-consumptive research?
  - How do we capture scholar knowledge to enhance data curation?