Digging into Metadata: Enhancing Social Science and Humanities Research

Significance

U.S. Team
College of Information Science and Technology, Drexel University, Philadelphia PA
PI: Dr. Mick Khoo (P.I.)
Team Members: Dr. Xia Lin and Dr. Eileen Abels

U.K. Team
MIMAS, University of Manchester, Manchester
PI: Diana Massam
Team Members: Hilary Jones, Andrew Priest, Anne Reed

University of Glamorgan
Dr. Doug Tudhope

Abstract

Metadata - 'data about data' - is a crucial component of information infrastructures that supports the finding, retrieving and sharing of resources; metadata in the social sciences and humanities helps researchers, educators, students and others to share data and information, and to collaborate in research, education and learning. Metadata work is highly resource intensive, and there can be significant technical problems with metadata quality. These issues are compounded by the fact that interoperability requires the use of a standardized metadata format; however, many collections of digital resources are described by local and specialized standards, and the metadata for these collections first has to be mapped and crosswalked to a standard format in order to be useful in cyberinfrastructure. Taken together, these constraints mean that there is a significant 'metadata gap' between the amount of digital resources requiring standardized description for use in cyberinfrastructure, and the number of people who can generate such descriptions. Machine-based alternatives to metadata generation show promise but can run into difficulties when faced with describing heterogeneous resources. Social classification techniques (tagging, folksonomies) show promise but are uncontrolled and often non-technical in nature.

We propose to develop a novel and innovative method for automatically generating tags for digital resources by digging into metadata records and digital resource content, and then performing text analyses of the existing metadata for each resource, the content of the resource itself, and any further descriptive information that can be associated with the resource. The technique will be automatic, scalable, and through tag clustering will permit resource discovery across multiple heterogeneous collections, without first having to crosswalk metadata to a standard format. This should represent an increase in productivity as well as a cost savings in terms of metadata generation.

The broader impact of this proposed work lies in its novel approach to the design, production and use of interoperable resource descriptions in information infrastructure. This approach would improve federated-like discovery across heterogeneous repositories for humanities and social science researchers. The results of the project would likely be extendable and applicable to many other small to medium multi-disciplinary metadata repositories that seek to add value for their users.

Digging into Metadata: Enhancing Social Science and Humanities Research

1. Introduction

The *iSchool* at Drexel along with our partners at the University of Manchester and the University of Glamorgan are pleased to present our proposal for Digging into Metadadata. The proposal addresses a significant data issue not covered in previous Digging into Data challenges: the technical and organizational issues of how to make sense of *multiple* heterogeneous datasets across computer networks for researchers in the social sciences and humanities (SSH). New large-scale information infrastructures are integrating vast amounts of data, information and knowledge across computer networks. Metadata - 'data about data' - is a crucial component that supports the finding, retrieving and sharing of resources across information infrastructures. Metadata in cyberinfrastructure helps researchers, educators, students and others to share data and information, and to collaborate in research, education and learning.

The work involved in creating metadata is complex, and despite advances in automatic metadata creation, is still largely a human and organizational practice. (Greenberg, 2010) Metadata work is highly resource intensive, and there can be significant technical problems with metadata quality. These issues are compounded by the fact that interoperability requires the use of a standardized metadata format, however, many collections of digital resources are described by local and specialized standards, and the metadata for these collections first has to be mapped and crosswalked to a standard format in order to be useful in cyberinfrastructure. (Khoo and Hall, 2010) Taken together, these constraints mean that there is a significant 'metadata gap' between the amount of digital resources requiring standardized description for use in cyberinfrastructure, and the number of people who can generate such descriptions.

We propose therefore to develop a novel and innovative method for automatically augmenting existing metadata sets to support interdisciplinary research in the social sciences and humanities (SSH). The method will generate tags for digital resources by digging into metadata records and digital resource content, and then performing text analyses of the existing metadata for each resource, the content of the resource itself, and any further descriptive information that can be associated with the resource. The analysis will generate tags that will be matched with subject categories in the Dewey Decimal Classification (DDC), a widely used multidiscipinary system. The technique will be automatic, scalable, and through tag clustering will permit resource discovery across multiple heterogeneous collections, without first having to crosswalk metadata to a standard format.

One of the goals of metadata standards and crosswalking is to support interoperability (Woodley, 2008) Metadata interoperability depends on crosswalking individual field to individual field. Crosswalking is complex, difficult and tedious, and it involves a range of complex organizational and technical work, and metadata crosswalk issues can stand in the way of interoperability. Without interoperability, however, researchers in the humanities and social sciences are hindered as they attempt to search across multiple datasets. If we are successful, we will provide a scalable approach to facilitate humanities scholars to search across large digital libraries, by building Dewey Decimal Classification-based tag indexes across different collections, and then clustering those tags. The method will support a form of federated search, but without having to generate standardized crosswalked metadata.

Figure 1 presents an overview of the proposed project (see also the Scenario of Use provided in the Appendix). In summary, viewing the figure from top down, the researchers, educators and students in the social sciences and humanities query the system, viewing the results in dynamic visualization interfaces based on topic maps. The digging analysis tool, which is transparent to the user, combines algorithms, latent semantic indexing, and clustering techniques to enhance the tags generated by the underlying DDC-based automatic tag recommender, PERTAINS. The information seekers are in effect accessing multiple linked repositories at one time, rather than searching each repository individually, much like a federated search.

Digging into Metadata: Enhancing Social Science and Humanities Research
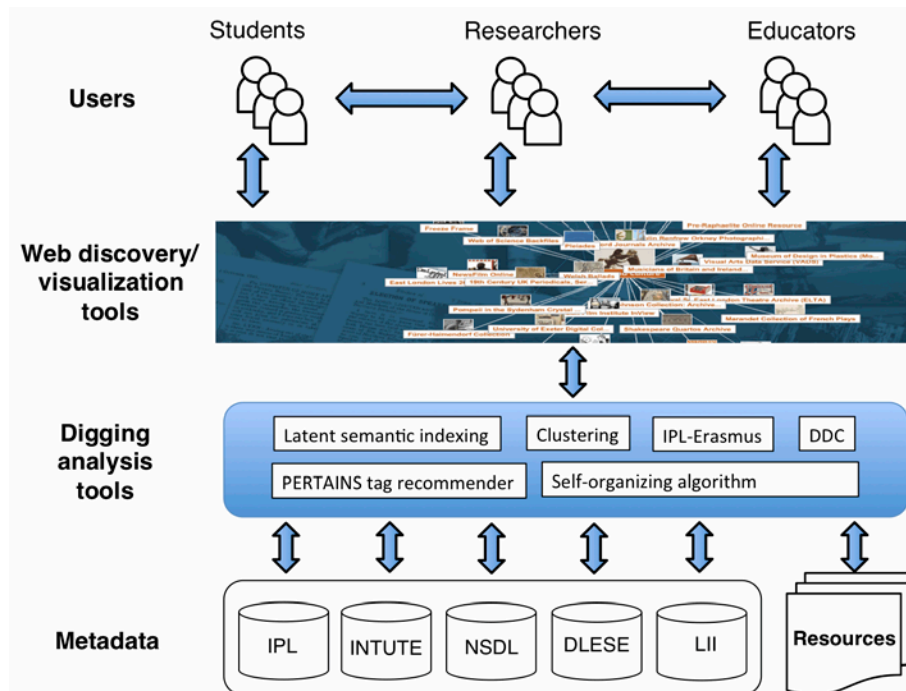

Figure 1. Overview of the Digging into Metadata Project

2. Research Questions

The metadata gap described above has restricted metadata production and has had negative consequences for data sharing in information infrastructures. There are a number of existing approaches to addressing this gap, including training more metadata specialists, automatic metadata extraction from resources, and training users to contribute metadata (Wilson, 2007). Each method has advantages and disadvantages (Greenberg, 2010). Our approach will adopt a unique approach by combining existing tools and methods developed by our individual partners in a novel configuration and approach that will support search across heterogeneous metadata without crosswalking and federated search.

- RQ1: How can we dig into metadata to improve discovery across heterogeneous repositories?
- RQ2: Do automatically generated tags from existing metadata records and resources increase user satisfaction in searches? If so, how?
- RQ3: How can automatically generated tags and user generated tags be clustered around standardized controlled vocabulary used in the metadata, and link uncontrolled to controlled vocabulary?
- RQ4: What are effective visualization techniques for presenting clusters of tags, metadata and search results?

3. Datasets

In order to improve effective use of metadata, and demonstrate applicability and scalability of the techniques, we will apply our proposed techniques to a heterogeneous set of metadata datasets. These datasets represent various domains and have been generated by different methods. To provide contrast, we have selected two repositories from the repository list provided on the DiD Challenge website, National Science Digital Library (NSDL) and Digital Library for Earth Sciences Education (DLESE), and

will also include two of our own repositories, ipl2 and Intute. This project will link our repositories (ipl2 and Intute) to NSDL and DLESE, without creating a formal mapping and crosswalk.

The *Internet Public Library (IPL)* was developed by Dr. Joe Janes and students at the University of Michigan's *iSchool* in 1995. Janes (1998) describes the history of the IPL. It is available throughout the world as a training tool for library and information science (LIS) programs as well as a freely available resource for the general public; it plays a major role in preparing a new generation of information professionals. In January 2007, the IPL was moved to Drexel's College of Information Science and Technology. In 2008, the *Librarians' Internet Index (LII)* was migrated to the *iSchool* at Drexel, with the intent to merge the two repositories and the name was changed to ipl2.   ipl2 includes selected high quality resources across twelve subject areas (including the social sciences and the arts and humanities) that are geared toward the general public and student users. The metadata records have been created by graduate student volunteers who receive some training and oversight but the repository is lacking consistent quality control. The LII included selected high quality resources with high quality metadata generated by librarians. As part of this merger process it was decided to crosswalk all existing IPL and LII metadata to Dublin Core. In the process, the implementation team encountered a number of 'metadata bottleneck' and associated issues. These issues were rooted in the original metadata formats being used, and the individual sociotechnical and organizational backgrounds of IPL and LII. (Khoo and Hall, 2010) Much work was required to understand these backgrounds in order to manipulate the metadata and share it.

Intute (http://www.intute.ac.uk/) was formed in 2007 by a merger of the eight former Resource Discovery Network (RDN) hubs. It is a collaborative project, based at Mimas, The University of Manchester, which provides a catalogue of high quality Web resources for students and researchers in UK higher education. Intute contains high quality metadata records created by information professionals and covers nineteen subject areas including the humanities and social sciences. In 2010, representatives of Mimas and the *iSchool* at Drexel began discussions about a possible merger of the ipl2 and Intute. Funding for Intute will cease in July 2011 and staff at Mimas are seeking alternative homes and uses for the data. Mimas is currently engaged in several other projects addressing these technical and organizational problems.

NSDL is a very large repository that provides organized access to high quality resources related to science, technology, engineering, and mathematics (STEM) K-12 and undergraduate education. NSDL provides a centralized metadata repository that contains mixed quality metadata harvested from various NSDL collections. For NSDL, metadata creation is a decentralized practice that includes both manual and automatic processing by partner projects and aggregated through automatic metadata harvesting and OAI-PMH.   In contrast, *DLESE* enforces a unified metadata creation procedure and involves information professionals and trained personnel to create metadata records. DLESE, a large repository, serves a more focused domain (earth systems) and maintains a relatively high quality metadata repository, relying on community members to create metadata tat was submitted via community tools, and quality controlled by metadata experts. By testing our methods on both NSDL and DLESE, the applicability of our proposed methods will be well established. [The DLESE metadata is also available through NSDL.]

As this project will utilize automatic algorithms to detect and establish semantic links among concepts and metadata records, having large datasets has many advantages. The algorithms we plan to use include clustering techniques, latent semantic indexing techniques, and other statistical-based learning and self-organizing algorithms. Working with large repositories improves the validity of the algorithms. Large repositories will also enhance the quality of recommended tags and linkages to be generated by the algorithms.

Table 1 summarizes features of the selected metadata repositories.

Repositories	Size	Metadata creation	Creation methods	Quality	Domains
IPL/LII	Small to medium	Volunteers	Supervised with training	Mixed	General and K-12 education
INTUTE	Medium Large	Information professionals	Decentralized	High	Multi-domains and comprehensive
NSDL	Very large	Harvested/semi-automatic	Decentralized	Mixed	STEM
DLESE	Medium Large	Information professionals	Centralized	High	Earth Systems

Table 1.   Features of the selected datasets

4. Approaches

4.1 User Studies

User-centered design is a key part of the project. Understanding social science and humanities (SSH) scholars' needs in the context of their everyday work practices, particularly the needs they encounter while searching across multiple and heterogeneous repositories, and then translating those needs into requirements and then tools, will be a central project activity. The Advisory Board therefore consists of local SSH researchers, librarians, and digital experts, will function as an 'at-hand' resource to provide access and insights into SSH users' needs. The Board will be able to assist the team in the initial identification of SSH researchers for information needs assessments, satisfaction studies, and usability testing during the project (for further details, see Work Plan, section 4.7).

4.2 Methods & Technologies for Digging Tool

This project proposes an integrated approach to improve the utilization of metadata through semantic integration of multiple metadata repositories. The proposed technology workflow is illustrated in Figure 2 below (the numbers in the figure correspond to the steps below).
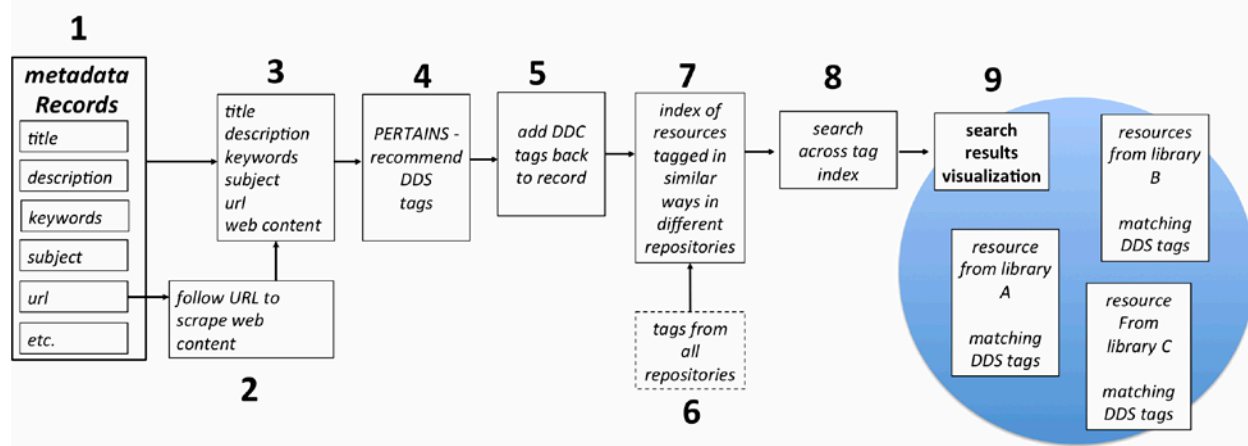

Figure 2 Technology Workflow

The steps are as follows:

1. For each resource in each repository that we are partnered with, combine the individual elements in the metadata record associated with that resource into a single document.

2. Identify and retrieve any resource content that can be reached through information in the catalog record (for instance by following links to html files, pdf files, etc.).

3. Combine the output of (1) and (2) above into a single document.

4. Carry out text analysis on the document created in (3) using the PERTAINS tool, in order to generate recommended Dewey Decimal System (DDC)-based tags to describe the resource.

5. Re-index the DDC tags back into the catalog record for the individual resource.

6. Repeat the process over all resources in all repositories in this project.

7. Use clustering algorithms to generate an index of similarity and match between resources in different repositories that have been analyzed and tagged by PERTAINS in similar ways.

8. When a user searches across the different repositories in the proposed project, they will be returned a list of resources in the different collections that have been tagged in similar ways.

9. The returned resources will be displayed in innovative discovery visualization interfaces.

4.3 Team Experience Related to Digging into Metadata

While both teams have been working on some DDC-based metadata enhancement projects independently, this project will provide the opportunity to bring the two efforts within the same framework.

Among the partners, we have been working on specific technologies to assess each of the approaches. The members of the ipl2 team have been grappling with metadata crosswalk and interoperability issues for over four years and the prospect of a merger with the Intute data has provided additional indicators of the importance of resolving these issues on behalf of researchers in all disciplines. Due to its complexity, current approaches to this kind of work appear not to be scalable, and alternate methods to achieve a crosswalk-like functionality and interoperability are being sought. In the past few months, work was initiated to use Dewey Decimal Classification (DDC) as a conceptual structure to unify two of our major collections, ipl2 and Librarians' Internet Index (LII). We envisage that DDC can be the "*tree*" while the localized "IPL subject headings" and "LII topic terms" could be the "*leaves*" attached to the tree to provide both standardized content structures and flexible local adoptions. The tree also becomes a bridge to link the two collections. Initial testing allows us to test the access of DDC through APIs provided by OCLC and some content analyzing tools that can be utilized for the attachment of "leaves" to "tree." If successful, much more "colorful leaves" (more collections) as well as user-generated and machine-generated tags can be attached to the tree to enhance the findability of metadata records and resources.

The UK partner team, including Mimas at the University of Manchester and the University of Glamorgan, created the PERsonalised TAgging Information Services (PERTAINS) and has used it to generate personalized, DDC-based tags for Intute resources. PERTAINS is an automatic tag recommender system based on selected controlled vocabularies or classification schemas. Binding et al. (2009) describe the details of PERTAINS in their final report to JISC. In this project, we will use PERTAINS as the main technology. We will enhance the PERTAINS tool with some well-tested content analysis algorithms such as concept clustering algorithms, latent semantic indexing, and self-organizing learning algorithms. Here,

Mimas staff at the University of Manchester have worked with the JISC Content tool, which visualizes JISC Collections digitized across all subject areas. This work included creating the subject taxonomy and devising the new database to support searching, browsing, searching/browsing by timeline, and an interactive mind ma (JISC, 2010). Faculty at the *iSchool* at Drexel have been working on prototypes of algorithms that are applicable to this project.

In addition, the *iSchool* at Drexel brings expertise in the area of human computer interaction and interface design. Since 2007, we have been working on three new interfaces for the ipl2 including a multi-window-based interface, a facet-based browsing interface, and a metadata-filter-based search interface.

In summary, our teams have been working on a series of technologies that represent different approaches to metadata processing and utilization. Based on these technologies, this project will create an integrated approach to dig into metadata repositories. We will use OCLC's APIs to get access to DDC, use PERTAINS to generate automatic tags, use Erasmus (the current tool being developed for ipl2) for manual editing and management of metadata, and use our innovative visual interfaces to test and experiment with new capabilities created as a result of digging into metadata.

4.4 Standards

All the repositories involved in the project use a form of Dublin Core metadata. The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description; its elements are broad and generic, usable for describing a wide range of resources. Each repository has also developed versions of 'qualified' Dublin Core, that is, versions of Dublin Core adapted to their particular subject domains and user communities.

It is recognized that standardized metadata supports interoperability between digital libraries. This project will enhance the standardized metadata. The Dewey Decimal Classification (DDC) is a standardized controlled vocabulary. The terminology used in DDC may be modified as part of this project to make it more user friendly for tagging. Advisory board members and SSH researchers will provide input into this effort.

4.5 Environmental scan

The different themes of our project are being addressed by other projects. However, we bring a new approach to current work found in the literature and through an environmental scan. Building an enhanced indexing to enable discovery service is a central theme of this project. In this regard, this project will create a discovery service similar to some publishers' discovery service such as Serials Solutions' Summon and EBSCO's Discovery Service (EDS). (Missingham, 2004) Elsevier has aggregated the content from Scirus, ScienceDirect, and Sopus into one single index (http://www.hub.sciverse.com). In this project, we will not only enhance discovery of metadata but also enhance the metadata by adding and recommending DDC-based tags to metadata.

Using knowledge organization systems to enhance tagging and metadata creation is another theme of the project. Spiteri (2010) summaries the trends and potentials of this approach. Matthews et al. (2010) provided empirical evidence to demonstrate the benefits of combining social tags and controlled terms to increase effectiveness of information discovery and retrieval. The PERTAINS tool, developed by partners of this project, is the first to generate standard-based tag recommenders from metadata. In this project, we would like to expand the use of PERTAINS to generate tags from both metadata and fulltext documents. We would also like to include user-generated tags.

Another theme of the project is to scrape websites to generate metadata automatically or semi-automatically. There are many open–source web scraping tools such as MIT's Crowbar (MIT-SIMILE,

2008) and Piggybank (MIT-SIMILE, 2010), that can be used for scraping keywords and their semantic relationships. In this project, our focus is to use web scraping to support automatic or semi-automatic metadata creation, which has gained increasing support by metadata specialists as reported in the study by Greenberg, Spurgin, & Crystal (2006).

Finally, this project will also advance research on using visualization interfaces for metadata-based discovery services. Two review papers summarize this approach. Shiri (2008) reported a study that examined 21 metadata-enhanced visual interfaces and concluded that visual interface was becoming increasingly used in support of user's information exploration and discovery. Lanzenberger, Sampson, & Rester (2010) discuss the usefulness of visual representation for semi-structured metadata.

4.6 Potential Risks and Risk-mitigation Strategies

Innovative approaches bring with them some risks. Theoretically, we should be able to generate clusters of resources from different digital libraries based on the tags generated automatically from those resources and their metadata records. This would enable a form of federated search without having standardized metadata. The fact that we are using an existing tool, PERTAINS, helps to mitigate the risks. We will use several approaches and "tweak" the PERTAINS recommender based on testing.

4.7 Work Plan

The project work plan and institutional responsibilities are described below and summarized in the Work Plan Gantt Chart that is included in the Appendix. The project is split into two main phases: Year 1 and Year 2. Year 1 work will focus on metadata augmentation through generating recommended DDC tags from existing metadata records. Year 2 will repeat the process, and will also include the generation of recommended DDC tags from the resource itself. The reason for this split is to allow the project to proceed to prototyping based on the relatively simple task in Year 1, and avoid an early dependency on the more complex task of retrieving the resource content prior to core research task of generating recommended DDC tags.

Year 1
In the first quarter, we will begin the project with an Advisory Board meeting. In addition to discussing a strategy for the overall project, we will focus on understanding information behaviors of SSH researchers. The Advisory Board will provide us with access to initial contacts in this community. Once these initial contacts in the SSH community have been established, they will be expanded through snowball sampling to identify a wider range of SSH scholars, researchers and students. Tool requirements will then be gathered continuously throughout the project through interviews, focus groups, and site visits.
During Q1, the catalogs of the individual repositories will be downloaded, and work will begin on submitting these to PERTAINS. As the PERTAINS analyses proceed, in the second and third quarters, the recommended DDC tags will then be added back to the relevant catalog records in each repository in a new record field. Following this, the research into aggregating these tags across repositories, and clustering and indexing them, will begin in the third quarter and continue in the fourth quarter, ending with test searches and evaluation across these aggregated tags and resources. Work will also begin on an interface prototype. The fourth quarter will include a follow up meeting with the Advisory Board for reactions to the results and preliminary interface testing.

Year 2
A similar research cycle will be carried out in Year 2, with the advantage that a prototype of the basic workflow, as well as a basic discovery interface, will already have been developed. In Year 2 therefore the research process will be elaborated to include more complex metadata augmentation sources, particularly the content of the resources being indexed, and more complex visualization techniques, including the topic map visualization. Usability testing will be conducted with SSH researches. Advisory

board meetings will be held in the first and fourth quarters.

Evaluation
In both Years 1 and Year 2, a range of evaluation activities will be carried out. We will have ongoing formative evaluation of the various interfaces and visualizations, using doctoral students trained in HCI methods working with ipl2. We will also have more formal summative evaluations for the three main components of the project, focused in the final quarters of Year 1 and Year 2: 1) retrieval effectiveness; 2) interface design; and 3) visualization of results. The summative retrieval evaluation will be carried out by the University of Glamorgan, and Mimas. Drexel will take the lead on the summative interface design and visualization evaluations.

5. Preparation of Graduate Students and Newer-to-the-Field Researchers

The results of this project will be integrated into the curriculum at the *iSchool* at Drexel, with the intent of making it available to other *iSchools* and Library and Information Science programs for their use. The *iSchool* at Drexel has extensive track record in the preparation of graduate students at both the master's and doctoral levels.   We have specific experience in integrating research related to the ipl2 into courses on digital libraries, collection development and reference services. Since migrating the IPL to Drexel in 2007, nearly 4,000 graduate students from library and information science programs across the country have taken advantage of the experiential learning experiences available. The various learning objects and syllabi created are available at a project website for a recently completed IMLS Grant at http://vll.ipl.org/learningobjects.html.

To address the merger of the IPL with LII, we formed a merger team that consists of faculty, staff, and doctoral students. Many of the doctoral students funded as IMLS fellows at the *iSchool* have been working with the ipl2 on a variety of research projects. These students have been actively engaged in publishing and presenting their work with their faculty mentors. [See profiles of our IMLS funded PhD students along with sample publications on our website: http://ischoolphd.org/index.php?option=com_content&view=section&layout=blog&id=7&Itemid=106]

6. Dissemination

The project team will develop a website that will provide information about the project and the progress being made.   We will incorporate a blog into the website and post updates on the blog. We will also use the social media tools already in place for the ipl2 to communicate results with interested members of the community.

We will share the tag suggestions generated for the various repositories so that they may in turn be shared with users of the repositories.   In addition, a database with the metadata will be made available to others for experimentation.

Additional dissemination will occur at major conferences including ASIST&T, Theory and Practice of Digital Libraries (TPDL), Joint Conference on Digital Libraries, and the iConference. Faculty on the project will develop and submit journal articles to appropriate publications, including the Journal of the American Society for Information Science and Technology (JASIST), and the International Journal on Digital Libraries.

References

Binding, C.; Priest, A.; Rigby, Janine; Tudhope, D.; and Williams C. (2009). PERsonalised TAgging Information Services (PERTAINS): integrating tag recommenders in UK national services. Final Report. Available online at: http://hypermedia.research.glam.ac.uk/kos/pertains/

Greenberg, J. (2010). Metadata and Digital Information. In *Encyclopedia of Library and Information Science, Third Edition*, 3610-3623. New York: Marcel Dekker, Inc

Greenberg, J.; Spurgin, K.; & Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. International Journal of Metadata, Semantics, and Ontologies. 1(1), 3-20.

Janes, J. (1998). The Internet Public Library: An Intellectual History. *Library Hi Tech*, *16*(2), 55-68.

JISC (2010). An introduction to our digital collections and archives. http://www.jisc-content.ac.uk/

Khoo, M., & C. Hall. (2010). Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. 10th ACM/IEEE Joint Conference on Digital Libraries, Brisbane, Australia, June 21-25, 2010, pp. 361-364.

Lanzenberger, M., Sampson, J. & Rester, M. (2010). Ontology visualization: tools and techniques for visual representation of semi-structured metadata. Journal of Universal Computer Science, 16(7), 1036-1054.

Matthews, B., Jones, C., Puzon, B., Moon, J., Tudhope, D., Golub, K. and Nielsen, M. L., (2010). An evaluation of enhancing social tagging with a knowledge organization system. *Aslib Proceedings*, 62 (4/5), pp. 447-465.

MIT-SIMILE (2008). Crowbar project. http://simile.mit.edu/wiki/Crowbar

MIT-SIMILE (2010). Piggybank project. http://simile.mit.edu/wiki/Piggy_Bank

Missingham, R. (2004). Reengineering a National Resource Discovery Service: MODS Down Under. D-Lib Magazine, 10(9), Sept. 2004.

Shiri, A. (2008). Metadata-enhanced visual interfaces to digital libraries. Journal of Information Science, 34(6), 763-775.

Spiteri, L. (2010). Incorporating facets into social tagging applications: an analysis of current trends. *Cataloging & classification quarterly*, *48*(1), 94-109.

Wilson, A. (2007). Toward Releasing the Metadata Bottleneck. Library Resources and Technical Services, 51(1), 16-28.

Woodley, Mary S. (2008). Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information. In Baca, M. (Ed.), *Introduction to Metadata.* Los Angeles, CA: The Getty Trust.