Final Report for Period: 07/2006 - 12/2006 **Submitted on:** 01/21/2007

Principal Investigator: McCue, Janet A. **Award ID:** 0437603

Organization: Cornell University State

Title:

SGER: Planning Information Infrastructure Through a New Library-Research Partnership

Project Participants

Senior Personnel

Name: McCue, Janet
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Lust, Barbara
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Crawford, Clifford
Worked for more than 160 Hours: No
Contribution to Project:

Name: Kedar, Yarden
Worked for more than 160 Hours: No
Contribution to Project:

Undergraduate Student

Name: Newman, Megan
Worked for more than 160 Hours: No
Contribution to Project:

Name: Fitch, Travis
Worked for more than 160 Hours: No
Contribution to Project:

Name: Hyland, Charles
Worked for more than 160 Hours: No
Contribution to Project:

Technician, Programmer

Name: Lowe, Brian
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Corson-Rikert, Jon

Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Pantle, Steven
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Webb, Frances
Worked for more than 160 Hours: No
Contribution to Project:

Name: Klein, Matthew
Worked for more than 160 Hours: No
Contribution to Project:

Other Participant

Name: Westbrooks, Elaine
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Steinhart, Gail
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Paulson, Joy
Worked for more than 160 Hours: No
Contribution to Project:

Name: Flynn, Suzanne
Worked for more than 160 Hours: No
Contribution to Project:

Name: Tobin, Theresa
Worked for more than 160 Hours: No
Contribution to Project:

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Robert Grotke (Cornell Laboratory of Ornithology)
Bill Arms, Claire Cardie, Carl Lagoze (Cornell computer science faculty)

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

See attached file.

Findings: (See PDF version submitted by PI at the end of the report)

See attached file.

Training and Development:

Students, graduate and undergraduate, have been engaged in this project in each of the research labs as well as in the Library itself.

In the CLAL, graduate and undergraduate students are being trained in the new data management techniques developed through this project, and a new section of a lab course has integrated data management and the DTA tool as a major component. These developments can help to transform the primary research process, so that the academic culture of future generations understands the importance of standardizing metadata originally in the research process, and begin to cultivate a new culture of data sharing and collaboration involving academic libraries as natural, effective partners.

The library has offered a training workshop on metadata creation to members of the USAEP project and participated in poster sessions for new students and researchers. The ability of the library to offer consultation and training, as well as potentially sufficient storage space for intermediate datasets, opens up new avenues for informal as well as formal collaborations.

Outreach Activities:

The several posters and papers at the series of conferences listed have provided continual outreach over the course of this SGER. (See Publications and Products.) In addition, professors Lust (Cornell) and Flynn (MIT) will teach a course in July 2007 at Stanford University (Linguistic Society of America summer institute) and offer a one day workshop on procedures of data management and research collaboration in a virtual environment; here the results of the SGER project will be shared.

Journal Publications

Books or Other One-time Publications

Lust, B., Westbrooks, E., Flynn, S. and Tobin, T., "Developing Adequate Documentation for Mutifaceted Cross Linguistic Language Data: First Language Acquisition.", ( ). article in book, in preparation
Editor(s): Furbee, L. and L. Grenoble.
Collection: Language Documentation: Theory, Practice and Values.
Bibliography: Publisher: John Benjamins.

Lust, B., Flynn, S., Blume, M. Corson-Rikert, J. and Lowe, B., "Searching Interoperability between Linguistic Coding and Ontologies for Language Description: Language Acquisition Data.", (2005). paper presented at workshop, Published
Bibliography: Electronic Metastructures for Endangered Languages Data (E-MELD) workshop, Cambridge, Mass. July 1-3, 2005. http://www.emeld.org/workshop/2005/papers/lust-paper.doc.

Steinhart, G.S., "Data Distribution and Archiving in Support of the Agricultural Ecosystems Program.", (2006). poster, Published
Bibliography: Presentation at the Annual Meeting Upstate New York Science Librarians. October 20, 2006. Binghamton, NY.

Lust, B., Westbrooks, E., Flynn, S. and Tobin, T., "Developing Adequate Documentation for Multi-faceted Cross Linguistic Language Acquisition Data.", (2005). poster, Published
Bibliography: Linguistic Society of America Conference on Language Documentation: Theory, Practice, and Values, MIT/Harvard, Cambridge, Mass.July 9-10, 2005.

Web/Internet Site

URL(s):

http://metadata.mannlib.cornell.edu/lilac/
Description:

Other Specific Products

Contributions

Contributions within Discipline:

Recent reports from both the National Science Foundation and the National Science Board outline issues related to digital data stewardship, and one of the persistent challenges identified is to understand the roles and responsibilities of various actors in the digital collections universe. The SGER team set out to analyze potential roles for the library by working with scientists in two distinct labs--one focused on linguistic data; the other on environmental science. Although the needs of the two research groups were different, our research allowed us to develop a conceptual starting point for expanding the research library's services in support of the needs of 21st century scientists.

The research scientists associated with this project state that this collaboration between the library and the laboratory is 'extremely promising' and that they are 'highly motivated' to work with the library staff. Both the technical expertise of the programming staff and the subject domain knowledge of the librarian are enhancing the ability of the research scientists to achieve their goals. The collaboration also facilitated efforts to share valuable research data with the broader community.

If the role of the library is to foster communication between scientists as Leibniz asserted, this project allowed the library to begin expanding its services to support an important aspect of scientific communication--research data. Not surprisingly, the efforts of the library were well received. Soon after the collaboration with the Upper Susquehanna Applied Ecology Program began, the laboratory included the library in its next grant application to the USDA; similarly, the library has been included in ongoing discussions across the campus that focused on research dataùfrom those held at the Theory Center to those conversations taking place in the Office of Research.

The focus on this grant was exploratory: to test whether research libraries might serve as natural partners in addressing the data management needs of the communities they serve. Our experiences shows that the conceptual model that we developed as well as the specific collaborations that we've fostered provide ample opportunities to expand the role and the services of the library in multiple ways--from metadata training and consultation to data dissemination and archiving.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Categories for which nothing is reported:

Organizational Partners

Any Journal

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development

Contributions: To Any Resources for Research and Education

Contributions: To Any Beyond Science and Engineering

Activities

The purpose of this Small Grant for Exploratory Research was to explore the issues surrounding a new type of collaboration between scientists and research libraries to support the preservation, discovery, and sharing of primary research data. With recent advances in computing and telecommunications technology, the stage is set for a major shift in the way science is conducted. Researchers and funding agencies are recognizing that data can be valuable for purposes beyond the studies in which they were originally collected, and some agencies are requiring data sharing plans as prerequisites for funding support. There is, however, a lack of established infrastructure to support the services necessary for handling research data. This grant investigated the premise that research libraries might serve as natural partners in addressing the data management needs of the communities they serve.

Initial work centered around child language acquisition data from the collection of the Cornell Language Acquisition Laboratory (CLAL) in the context of the formation of a Virtual Center for Language Acquisition (VCLA). This collaboration prompted consideration of the full range of services a library might provide, from preservation reformatting of legacy analog multimedia resources to high-level metadata exposure and sharing of completed projects.

The project team set out to assess whether or when such collaboration is feasible or desirable, what significant challenges or costs are associated with these activities, how to balance the responsibilities of laboratory and library, and to develop a conceptual model for collaboration with an interest in promoting the preservation and discovery of resources valuable for interdisciplinary research.

Activities with the Cornell Language Acquisition Laboratory / Virtual Center for Language Acquisition

The collaboration with the Cornell Language Acquisition Lab and its related data prompted consideration of fundamental questions. What constitutes the data? Where is the boundary between data and metadata? Unlike other sciences where analysis may be performed directly on observational data or measurement values, the CLAL's analytical work is predicated on an intermediate stage of transcription from language recordings of audio speech streams in both naturalistic and experimental environments to text-based representations of the linguistically meaningful segments of the utterances conveyed therein. This process involves the judgment of a trained researcher, and multiple alternative transcriptions are possible for a given speech stream. Although the transcribed speech constitutes the "data" in the context of a particular research project, another researcher might require access to the original audio waveforms in order to confirm findings or to perform an alternative analysis. Thus it was clear from early on that the successful preservation and dissemination of the CLAL's data would involve the management of relationships: between original media files, metadata about the recording sessions, transcriptions of the speech, further segmentation and analysis, and published materials describing experimental results.

Although the library was able to outline best practices for the digitization of the audio recordings and plan for long-term storage of the resulting files, the close coupling between transcription metadata management and the linguistic research process suggested that this activity be conducted outside the library, in the domain-specific research lab. To this end, the CLAL developed a Data Transcription and Analysis (DTA) software tool for management of recording session metadata and the related transcription data. This tool is intended to provide a standardized conduit between the original primary data collection and analyses conducted by the research lab, and the systematic and structured metadata representation required for widespread dissemination and discovery procedures, such as fostered by a university library. Shared use of the DTA tool would also serve as a primary component of a Virtual Center for Language Acquisition (VCLA), comprised of researchers across multiple institutions and disciplines sharing a focus on or interest in language acquisition research.

The SGER project team thus began to conceptualize the CLAL-library partnership as a layered system wherein the library's key role would be to provide a top-level discovery mechanism using high-level metadata about research activities, as well as to provide the necessary links to allow one to "drill down" to the level of the primary audio and video recordings made available through research labs under appropriate access restrictions.

In summary, collaboration with the CLAL included:

- investigating the costs of and documenting best practices for digitization and long-term preservation of the CLAL's extensive audio collection;
- developing a conceptual model for research data and metadata management assuming multiple levels of interest and control;
- consulting on metadata standards for development of the research lab DTA tool;
- developing ways to generate Open Language Archives Community (OLAC) metadata from data stored in the CLAL's Data Transcription and Analysis (DTA) database;
- developing tools to generate basic OLAC metadata for continuously-growing new projects;
- monitoring developments in discipline-specific ontology initiatives such as GOLD (General Ontology for Linguistic Description) for applicability to language acquisition data and the DTA tool;
- developing an ontology-driven public web portal (http://vcla.mannlib.cornell.edu/) to provide information about CLAL and VCLA participants as well as a future means of access to data;
- participating in a teleconference organized by the VCLA, involving University IRB members and representatives from federal granting agencies (NSF and NIH), on initial development of inter-university standards for human subjects protections necessary to the management and dissemination of human subject data.

MIT collaboration

The SGER project also initiated collaboration between two research libraries in order to explore opportunities and requirements for systematic research data discovery and dissemination through university libraries in general. The MIT library was chosen as a logical partner for this exploration, given its role in coordinating the DSpace Federation and the OpenCourseWare project.

The Cornell-based project team met with MIT collaborators on two occasions (December 2004 and July 2005). The principal MIT collaborators were Professor Suzanne Flynn and Librarian Theresa Tobin, with MIT metadata librarians and members of the SIMILE (http://simile.mit.edu) and Open Courseware (http://ocw.mit.edu/index.html) projects participating in certain meetings.

Discussions primarily focused on the foundations for inter-library exchange of research data and materials. Specific topics included the evolution of the institutional repository as represented by current installations at MIT and Cornell, the challenges in developing financial models across subject domains and different institutions, the development of policies and support for populating institutional repositories with research data, new data grid developments, patterns and potential for metadata librarian involvement in research projects on a consulting basis, and the necessity to calibrate metadata standards across libraries and labs.

Activities with the Upper Susquehanna River Basin-Agricultural Ecology Program

A supplement to the original SGER grant award enabled engagement with a second research group conducting ecological research on the Upper Susquehanna River Basin. This work enabled the project team to consider a significantly different scenario for library handling of research data in a research field offering an existing repository structure and a higher percentage of "born-digital" data, to evaluate how the results of the initial project could be generalized across domains.

The Upper Susquehanna River Basin Agricultural Ecology Program (USAEP) at Cornell University aims to better understand the sources and sinks of nutrients and sediments in the New York portion of the Susquehanna watershed. In addition to collecting data and developing numeric and spatial models, the research group has access to at least thirty years' worth of data for some of its research sites. Such long-term records are of significant value to environmental scientists, but this potential will be lost if researchers do not have the tools and training to develop high-quality metadata for discovery and evaluation of these datasets. Participants are highly motivated to make their data publicly available by providing online access to benefit other scientists, policy makers and managers, and the public. To that end, the collaboration with Mann Library included:

- identifying a discipline-specific metadata standard, Ecological Metadata Language (EML);
- reformatting and documenting existing long-term datasets;
- training project collaborators in the use of existing metadata creation tools to create documentation for their datasets;

- submitting research datasets and associated metadata to Cornell's institutional digital repository (DSpace);
- developing tools to produce DSpace metadata from EML metadata;
- providing a wiki for the group to use as a collaborative tool;
- creating a public web portal (http://usaep.mannlib.cornell.edu/) to provide information about the project and participants, as well as a future means of access to project datasets.

References

Berkley, C., Jones, M., Bojilova, J., & Higgins, D. (2001). Metacat: A schema-independent XML database system. 13th International Conference on Scientific and Statistical Database Management (SSDBM 2001), Jul 18-20 2001, 171-179. Online: http://dx.doi.org/10.1109/SSDM.2001.938549. Retrieved 11/27/2006.

Ecological Metadata Language (EML). Online: http://knb.ecoinformatics.org/software/eml/. Retrieved 11/27/2006.

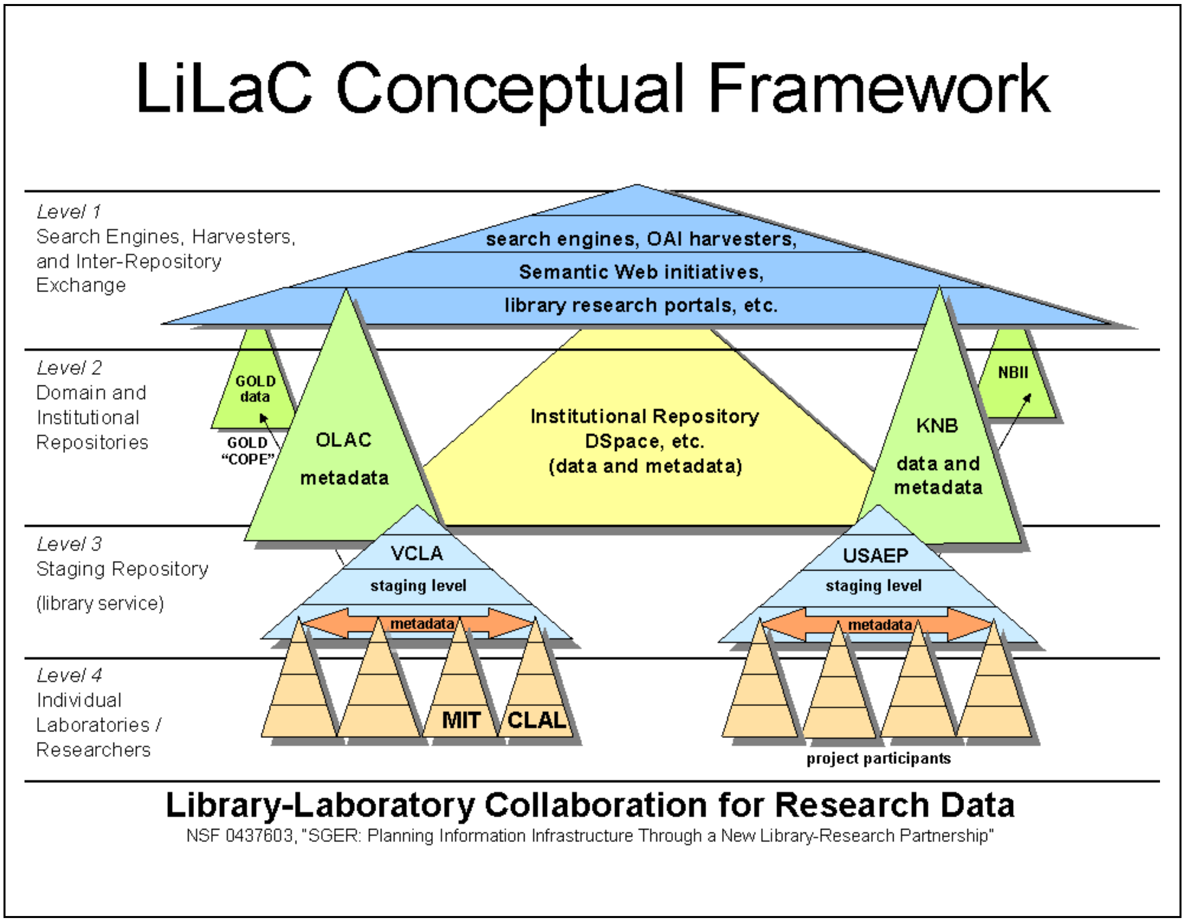
Higgins, D., Berkley, C., & Jones, M. B. (2002). Managing heterogeneous ecological data using Morpho. Proceedings of 14th International Conference on Scientific and Statistical Database Management, 24-26 July 2002, 69-76. Online: http://dx.doi.org/10.1109/SSDM.2002.1029707. Retrieved 11/27/2006.

Findings

The overall findings from our original SGER project activities and the supplement period include:

- A conceptual model for developing and disseminating metadata and data from individual discipline-specific research labs with university library-led data dissemination and discovery procedures. The model that was developed through work with the Cornell Language Acquisition Laboratory, and revised and expanded through additional work with the Upper Susquehanna Agricultural Ecology Program, has as its primary goal the creation of broadly-based, high level discovery opportunities for widely differing data from a variety of disciplines, institutions and laboratories. This requires the acknowledgement that disciplines do not organize or understand their data in mutually compatible frameworks, a factor unlikely to change. The innovative nature of research means that any given project may create new concepts or ontologies to describe its work. This is a vital aspect of invention even though the highly variable result may be difficult to incorporate into a unified system.

  The hierarchical model for preparing and disseminating metadata attempts to work within this framework by creating vertical bridges between detailed lab or research-area data description practices and the more standardized metadata accepted by disciplinary or institutional repositories. The image of this model below uses interlocking triangles to show how each successive layer provides exposure on a wider scale through generalization, aggregation, and indexing into broader search engines. Text labels within the colored shapes indicate implementations in the context of our two research labs, including avenues for future development.


LiLaC Conceptual Framework
Level 1
Search Engines, Harvesters, and Inter-Repository Exchange
search engines, OAI harvesters,
Semantic Web initiatives,
library research portals, etc.
Level 2
Domain and Institutional Repositories
GOLD data
GOLD "COPE"
OLAC metadata
Institutional Repository DSpace, etc. (data and metadata)
KNB data and metadata
NBII
Level 3
Staging Repository
(library service)
VCLA
staging level
metadata
USAEP
staging level
metadata
Level 4
Individual Laboratories / Researchers
MIT
CLAL
project participants
Library-Laboratory Collaboration for Research Data
NSF 0437603, "SGER: Planning Information Infrastructure Through a New Library-Research Partnership"

- Identification of typical steps involved in the process of converting from individual discipline-specific primary research metadata (however loosely defined in practice) to metadata records acceptable for ingest into disciplinary and institutional repositories, along with associated access restrictions or special technical requirements (e.g., ability to process spatial data).
- Initial demarcation of where discipline-specific research labs must develop and impose data normalization and/or transformation procedures sufficient to prepare data for re-use.
- Identification of functional requirements for and initial development of tools and procedures required for managing metadata as semantic links between resources, recognizing a need to develop local ontologies or reuse/extend specialized ontologies to model the research contexts of particular labs which then link to semantic representations of metadata in established schemas.
- Verification of the applicability of the conceptual model and associated data and metadata preparation procedures across widely differing research labs (one language based and involving human subjects, and one ecologically based).
- Initial estimation of costs associated with component steps of the data documentation process, expressed as levels of effort and resources associated with data conversion, metadata creation, and development of submission packages for appropriate repositories. Based on our experience, cost models can vary widely; some economic analysis has been done, but more is needed. The interim SGER report[1] lists some estimates of costs, largely associated with the conversion of analog media to digital files. The costs associated with the USAEP collaboration were much more modest due to the born-digital nature of the data and the use of existing software products freely available through the Knowledge Network for Biocomplexity (KNB). To better realize this project's conceptual model of metadata management will require the development of new software tools; as these become available, the costs associated with their implementation and use will be easier to forecast. The cost of long-term preservation of the data itself is naturally dependent on the size and format of the specific datasets and on the business models of the target institutional or domain repositories. MIT Libraries' article on the business model for DSpace is an example of a relevant analysis.[2]

In summary, in order to develop a research lab-university library infrastructure, this project has begun to leverage the library's existing expertise in preservation, archiving, and metadata creation, building on the existing ontology-software tools the library and the research labs have developed, and introducing a new conceptual framework that divides the task of data sharing into discrete levels that can be managed and presented in different ways not only for different audiences but respecting political divisions and control issues that will always be present throughout the laboratories and institutions of academia.

These general overall project findings were supplemented by discipline-specific results achieved by the two research labs through their collaboration with the University Library on this project.

[1] http://hdl.handle.net/1813/5240

[2] Barton. M. R., and Walker, J. H. (2003). Building a Business Plan for DSpace, MIT Libraries' Digital Institutional Repository. Journal of Digital Information 4:2. Online: http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Barton/. Retrieved 01/15/2007.

Results of CLAL activities

Through collaboration with the University Library in this SGER, and in pursuit of the ultimate aim of this SGER project, (i.e., building an infrastructure for data discovery), CLAL has been able to attain a number of developments including the following:

Primary data management and preservation

- Inventoried and organized physical data (audiotape) holdings in the CLAL (More than 4000 tapes and related written and electronic records).
- Developed procedures for data preservation, including digitization and labeling, to promote high-quality reformatting of original recordings in the CLAL's digitization efforts.
- Applied for funding to digitize additional primary research data. A pending proposal was submitted to the National Endowment for the Humanities on July 25, 2006: "Language Acquisition Digital Archiving: the South Asian Component." Joy Paulson and Barbara Lust, co-PIs.

Data description

- Implemented a beta version of a data management tool (Digital Transcription and Analysis tool) capable of documenting administrative, technical, observational, and interpretive components of language acquisition data and exporting the resulting datasets and associated metadata.
- Developed new data management techniques that allow bridging between primary metadata collections and disciplinary or library standards (in this case, OLAC) for data dissemination and discovery.

Data dissemination and access requirements

- Defined requirements for tools to allow interdisciplinary access to primary and secondary data, and potentially, the intermediary levels of data conversion.
- Developed websites to demonstrate metadata integration and opportunities for cross-disciplinary discovery.

Results of USAEP activities

During the supplement period of the project, the conceptual framework was further tested through application to the USAEP project. The collaboration has made good progress documenting the USAEP group's historical data, working with the responsible researcher to collect information on site locations and research methods. In this process we have employed software tools developed by the National Center for Ecological Analysis and Synthesis in conjunction with the Knowledge Network for Biocomplexity. The USAEP group has benefited from a workshop offered by the library on creating EML metadata with the Morpho tool (Higgins et al. 2002), and the library has set up two local metadata databases using the Metacat XML database system (Berkley, Jones et al. 2001): one for training, and one for completed metadata records. These activities have sparked interest among some of the researchers in extending these services to other group research projects, as well as to individual researchers.

Some additional requirements and challenges have also become evident. By relying on specific, existing tools available in the ecological discipline, we've committed researchers to working within the confines of those tools and standards, except where relevant crosswalks have been developed. A more flexible approach that allows for ex post facto decision-making regarding data publication is desirable. For example, the group has a significant amount of geospatial data, which might be appropriately published to the Knowledge Network for Biocomplexity (which utilizes EML), as well as a local or state-level geospatial data repository (most of which require the FGDC metadata standard for geospatial data). While an EML to FGDC crosswalk does exist, this may not always be the case when a researcher wishes to publish data to multiple domain repositories with different standards.

The group's need to share preliminary data also has identified further requirements for our conceptual model of library-laboratory collaboration. Currently, researchers have two options to share data within the group: attach data files to wiki pages, and upload data files to a local installation of Metacat. The limitations of attaching files to the wiki are that it is not possible to establish relationships between data and metadata in that environment or to exploit the content of metadata records. In addition, the wiki was never intended as a distribution platform for data sets. Researchers in this group also find the Morpho-Metacat interface somewhat awkward to use in terms of reviewing available data, and while we could develop a more user-friendly web-based means of access to the same content, a unified set of tools available through a single interface would go a long way towards addressing these problems in a more general way. These issues informed our thinking for a recently submitted proposal (to NSF) to develop a more generic, data-staging repository to facilitate the documentation and transmission of research data sets from a variety of disciplines to domain-specific repositories and/or institutional repositories.