Data Curation and Distribution in Support of Cornell University's Upper Susquehanna Agricultural Ecology Program

Gail S. Steinhart*
Albert R. Mann Library
Cornell University
Ithaca, NY 14853
GSS1@cornell.edu

Brian J. Lowe
Albert R. Mann Library
Cornell University
Ithaca, NY 14853
BJL23@cornell.edu

Abstract

Effective documentation, curation, and provision of access to scientific data are essential to derive the full benefit of research data, both for participants in specific research projects and for the entire scientific community. Academic research libraries are positioned to be important partners in such endeavors, although success will depend in part on expanding and changing the customary roles of, and relationships between, researchers and libraries. Cornell University's Albert R. Mann Library is collaborating with the Upper Susquehanna Agricultural Ecology Program at Cornell to document and distribute the group's research data. In addition to collecting data and developing numeric and spatial models, the research group has access to approximately thirty years worth of observational data for their research sites, which are of significant value to environmental scientists. The approach includes identifying and using discipline-specific metadata standards in order to facilitate participation in discipline-specific data and metadata sharing initiatives, at the discretion of individual researchers. Training is provided for project collaborators in the use of existing metadata creation tools to create documentation for their datasets. "Pre-publication" data and metadata are stored in a database accessible only by project members, to facilitate early sharing and collaboration within the group. Complete, documented data sets and complete metadata records will then be deposited in Cornell's DSpace installation. As a test case, the historic data sets are being formatted and documented for deposit in DSpace. A public web portal provides information about the project and participants, as well as a future means of access to project datasets.

* *Corresponding author*

Introduction

The issue of curation of scientific research data is receiving a good deal of attention. Increasing computational capacity and new tools, coupled with the accelerating accumulation of data in many disciplines, are giving rise to new modes of conducting research. Researchers increasingly make use of data they themselves did not collect. Funding agencies are beginning to recognize the need to document and archive the data that result from funded research, and in some cases suggest or require researchers to do so (e.g. Division of Ocean Sciences, 2003). Also significant is a recommendation of the Association of Research Libraries that the NSF develop data sharing policies and require data management plans as part of every research proposal (ARL Workshop on New Collaborative Relationships, 2006).

A central issue with respect to data curation is whether data repositories are best managed by organizations affiliated with specific disciplines, or whether they should be managed institution by institution (Messerschmitt, 2003). Given the significant investment in developing domain-specific standards and tools for the description, management, and use of data, it makes little sense for generalist institutions such as libraries to attempt to replicate those tools and services, if and where they exist. Organizations hosting domain repositories are best suited to manage by proxy, for their designated communities, issues such as technical standards and processes, ontology development, collection management, peer review, and other issues of interest to specific communities (Consultative Committee for Space Data Systems, 2002). However, infrastructure to promote and support the curation of digital data is subject to uneven development and availability across disciplines, institutions, and research scales. Some so-called "big-science" projects have relatively well-supported data infrastructures, while independent researchers, or researchers collaborating on smaller projects, may not receive the support they require to curate their own data and make it available (Carlson, 2006).

This uneven development of infrastructure and support, and the need for assistance to researchers suggests an important role for an institutionally based actor in this arena. We suggest that research libraries may effectively fill that role by providing services and infrastructure to support activities related to data curation early in the research life cycle, with the ultimate goal of facilitating the transfer of completed data sets and metadata to domain-specific and/or institutional repositories. Researchers may need guidance on how to prepare and describe their data, and guidance on where to submit completed data sets. Specific areas where support may be lacking include identifying an appropriate domain-specific or institutional repository for data submission, formatting data, describing data using appropriate tools and standards, and completing the submission process itself. Reports of lessons learned from various data management initiatives for specific projects reveal the need for assistance to researchers, and it is widely asserted that local support for data creators in the curation process is an important key to the success of such activities (Glover et al., 2006; Karasti et al., 2006; Lord & Macdonald, 2003).

Success in this area will depend at least in part on addressing changes in customary roles for both research communities and libraries. For many data creators, the idea and practice of documenting data and preparing it for public dissemination or archival is a significant shift in research culture, requires tools and skills they may not have, and may be seen as an additional burden in terms of time and resources or as an infringement on their research. From the perspective of research libraries, this process requires new types of partnerships with researchers in their home institutions, as well as with those engaged in curation, preservation, and distribution efforts in other arenas.

While these social and cultural issues are significant, there are also technical challenges that will not be addressed in this paper. Standards, best practices, and infrastructure continue to evolve, so strategies must be flexible and adaptable, yet dependable. Repositories must employ appropriate digital preservation practices, for example as outlined in the RLG/NARA audit checklist for trusted digital repositories (RLG-NARA Task Force on Digital Repository Certification, 2005). Finally, the computing infrastructure required may be significant, and partnerships with campus IT organizations may be beneficial. In spite of these challenges, the potential benefits to libraries for engaging in data curation are substantial. Locally, these include developing new partnerships between libraries and their constituents. More broadly, by participating in current efforts to develop a data curation infrastructure, libraries contribute to the discussion and development of scholarly communication practices in the broadest sense (Messerschmitt, 2003).

Scientific Context

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 upper (or New York) portion of the Susquehanna River watershed. The Susquehanna River is the largest tributary of Chesapeake Bay, and the single largest source of nutrients to the bay. New York State is committed to reduce the impact of its part of the Susquehanna River watershed on the Bay. An improved understanding of the sources and movement of nutrients and sediments in the Upper Susquehanna River watershed can inform management decisions aimed at reducing nutrient and sediment inputs to the system, and rural landscapes in general.

The research effort is interdisciplinary and inter-institutional. The 30-40 participants include graduate students, faculty, and staff from six different departments at Cornell University, and five additional institutions and organizations (Cornell Cooperative Extension of Chemung County, Institute of Ecosystem Studies, University Maryland Center for Environmental Science, University of Nebraska-Lincoln School of Natural Resources, Upper Susquehanna Coalition). Research funding is provided by the United States Department of Agriculture Cooperative State Research, Education, and Extension Service. Principal investigators redirect funding to project participants via competitive grants to students, faculty, and research associates. Recipients of funding are expected to participate in project-wide events, and share their results and data. Collaborators from Mann Library attend these events, meet with project principal investigators and individual researchers, and communicate with the entire research team as needed.

The research group has articulated two goals that are supported by this collaborative data curation and distribution effort. First, the project should serve as a source of valuable data and insight for people in county, state, and the federal government, NGOs, as well as members of the public seeking to improve water quality in the Upper Susquehanna Basin and the Chesapeake Bay. This will be accomplished by documenting research data, making it available online, and linking it to a public portal on the web that provides access to research data and general project information. Second, the project aims to serve as a catalyst for innovative cross-disciplinary research at Cornell that will produce better scientific understanding of nitrogen, phosphorus, and sediment cycling in the Upper Susquehanna basin and the Chesapeake Bay, by providing an easy way for researchers to know what other project participants have already done, are doing, and are planning to do. To meet this goal, the group is using tools that facilitate communication and sharing of data and documents within the group, privately, prior to publication. Collectively, these goals illustrate the group's commitment to collaboration, and because of the applied nature of the research questions, sharing their results and data publicly.

Data Curation Strategy

We are experimenting with a model where research libraries provide local support and local leadership to propagate metadata (and in some cases, data) "up" to broader and more visible repositories, which in turn expose content to discovery mechanisms such as metadata harvesters and search engines (Figure 1). While the figure emphasizes the movement of metadata and data among repositories, our approach also includes services to researchers to facilitate the transmission of data and metadata to discipline-based and institutional repositories, and involvement by library staff early in the research cycle, much in the same manner described by Green and Guttman (Green & Gutmann, 2006).

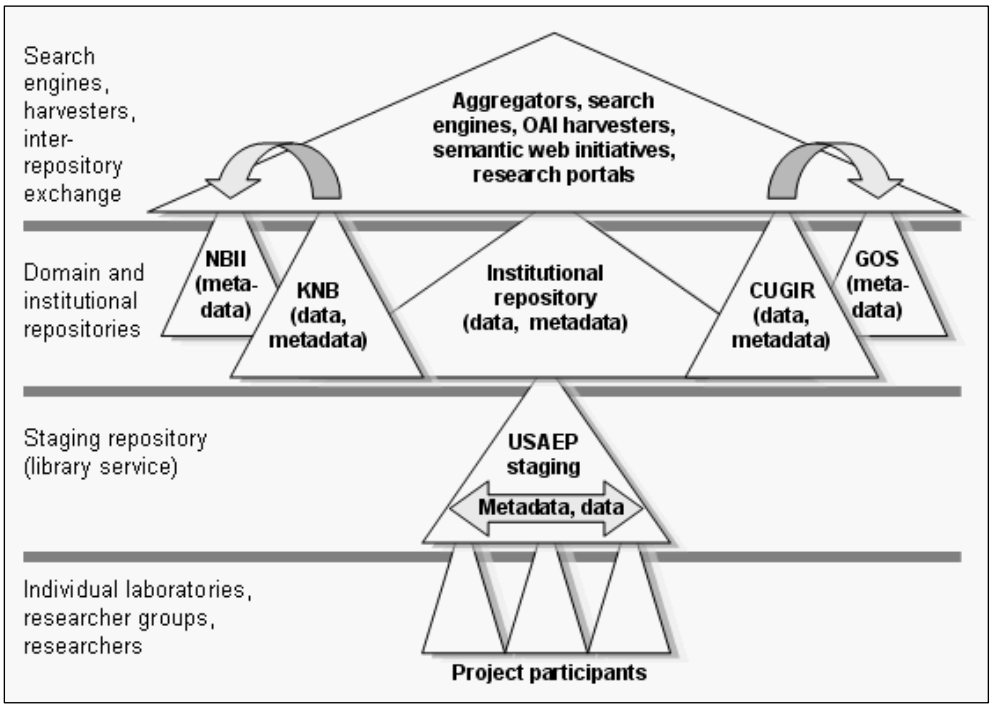

Figure 1. Conceptual model for the creation and movement of metadata and data from individuals and research groups to systems supporting sharing with collaborators, and eventually with the public. The "staging" area represents local infrastructure to support sharing within a group of researchers. When data and metadata are ready for public release, they may be submitted to an institutional repository and/or discipline-specific repositories, which may in turn expose their content for harvesting by other repositories. In this particular example, the National Biological Information Infrastructure (NBII) harvests metadata submitted to the Knowledge Network for Biocomplexity (KNB), and Geospatial One Stop (GOS) harvests metadata from the Cornell Geospatial Information Repository (CUGIR). Institutional repositories may be indexed by web search engines.

Because this effort, which represents an expanded role for the library, also involves processes that are new to most participating researchers, we've adopted the view that our efforts must support the day-to-day work of scientists and be consistent with their goals, or there will be little incentive to participate. In some circumstances we've provided a higher level of service than we might choose to sustain over the long term, to make barriers to participation as low as possible and foster a culture and appreciation of data curation among researchers. With these principles in mind, we planned a collaboration with the USAEP group that includes a preliminary survey of data collection plans, identification and selection of relevant standards and tools, and training and guidance on data and metadata preparation for researchers, and identification of suitable institutional and discipline-based repositories for the final deposition of data and metadata.

Planned data collection efforts

The first generation of project grants was awarded in spring of 2006. Some participants began collecting data actively almost immediately, or continued previous, related data collection efforts, while others planned to begin field data collection in 2007. Field data collection efforts are diverse in topic and include both experimental and observational data. Examples of experimental data collection include the effects of willow char amendments on agricultural soils, wetland plant species responses to changes in sulfur and phosphorus cycling, changes in groundwater chemistry as a result of chemical amendments, nitrous oxide fluxes and nitrogen leaching in soils under different cropping systems during winter freeze-thaw events and manipulations of snow cover, and changes in forest and old field soil chemistry as a result of nitrogen fertilization. Examples of observational data collection include measurements of dry atmospheric nitrogen deposition, stream and groundwater chemistry, stream stage and flow, air and water temperature, water table elevation, physical and chemical characteristics of soil, plant tissue chemistry, and cesium-137 in stream sediments. In addition, some participants are engaged in adapting simulation models to this particular research context, and will produce different research products from those researchers who are actively collecting data.

All research group participants who provided preliminary information on their data collection plans described their data sets as incomplete and in progress. A few reported that they plan to continue collecting data beyond the end of the funding period of this particular project. Most report Microsoft Excel as their working file format, and are not yet able to estimate the final volume of data they will collect. Additional types of data reported by participants include SAS files, geographic information system (GIS) files, comma-delimited text, and digital photographs. Those engaged in simulation modeling may produce executable files, model input and output files, and complex Excel spreadsheets that include formulas and macros. This preliminary information suggests that the majority of the data produced by project participants will be tabular and readily reformatted to a non-proprietary format such as comma-delimited text files, but there will be some exceptions.

In addition to the planned data collection activities described above, the research group also has approximately 30 years of observational data collected by an emeritus professor at Cornell. These data include stream and ground water chemistry and related measurements from sites of interest to the research group, and are of interest to the group both for immediate use, and as an historic body of data of likely use to future researchers. We are currently working with the data owner to collate, format and document these data, an experience that has presented us with some interesting challenges.

Metadata and data management

Metadata serve as documentation for data, describing the content, purpose, structure, format, and accessibility of datasets. Interdisciplinary and collaborative science creates an important demand for a set of "instructions" for researchers to make sensible judgments about whether and how they might use data provided by their colleagues. Metadata also serve a functional role in digital repositories, providing the raw material that makes it possible to display information about a dataset, and for search engines to index repositories and deliver results to users.

In considering choices for metadata standards to adopt for USAEP data, we had two primary options. The Biological Data Profile (BDP) for the Federal Geographic Data Committee (FGDC)

Content Standard for Digital Geospatial Metadata (CSDGM) includes additional fields for biological information (FGDC Biological Data Working Group, 1999). The CSDGM standard is well-established for geospatial metadata, and the National Biological Information Infrastructure has adopted the CSDGM with BDP as its standard for biological data (Strategic plan for the USGS national biological information infrastructure, 2005). Ecological Metadata Language (EML) was developed at the National Center for Ecological Analysis and Synthesis (NCEAS), specifically for the ecological sciences. While it accommodates geographic information in at least as much detail as the CSDGM-BDP, it is more robust in its capabilities to handle information of interest to ecologists, including detailed description of tabular data. In addition, EML is a modular, XML-based standard, which makes it more flexible and extensible (Fegraus, et al., 2005). Finally, an easy-to-use, platform-independent metadata editing application (Morpho) is available. Morpho also allows users to interact with Metacat (also developed at NCEAS; Jones et al., 2001), a metadata database, to upload metadata and data to a server, and to search, view and retrieve data and metadata on the system. Mann Library hosted an EML and Morpho training workshop, and offers individual consultations to researchers on an as-needed basis. For researchers wishing to document geospatial data with CSDGM standard, and to submit it for distribution to a geospatial data repository, assistance in preparing metadata is also available.

For long-term storage of data, and as one option for public distribution, we established a collection for the research group within Cornell's institutional repository (based on MIT's DSpace). Our main reason for doing so was expediency. Cornell already had a functioning DSpace installation; its ease of use and the institutional commitment to maintain its contents made it a practical choice. It is our plan to have researchers submit both data and metadata to the library, to automatically extract the information required to populate a DSpace metadata record from each detailed metadata record, and deposit both the data and detailed metadata record in the DSpace collection.

Because DSpace's primary function is one of storage and online access, and not one of communication of project information to the public, we also created a public web portal (http://www.usaep.mannlib.cornell.edu/). It is modeled after existing portals created at Mann Library for the Life Sciences community at Cornell (VIVO, the virtual life sciences library: http://vivo.library.cornell.edu/), and extended to other individual colleges at Cornell as well as the physical and social sciences, with support from the Provost (Caruso et al., 2006). The portal presents information on the project in context, and will include links to data sets deposited in DSpace. Information can be added to the portal using a web interface, making it possible for designated project participants to update content in the portal.

Finally, publication of metadata and data to discipline-specific initiatives is an option that individual researchers may choose to exercise. The use of EML for project metadata makes it very easy for researchers to publish their data to the Knowledge Network for Biocomplexity (http://knb.ecoinformatics.org/), which in turn automatically converts EML metadata to CSDGM-BDP and exposes it for harvesting by the National Biological Information Infrastructure (http://nbii.gov/). Creators of geospatial data choosing to create their metadata as CSDGM documents have the option of participating in various geospatial data initiatives, including the Cornell University Geospatial Information Repository (http://cugir.mannlib.cornell.edu/), also operated by Mann Library. CUGIR metadata are regularly harvested by Geospatial Onestop.

Supporting the collaborative research process

As mentioned earlier, we view support of the day-to-day research process as essential to encouraging participation by researchers in the work of data curation. In addition, one of the groups' stated goals is to serve as a model for interdisciplinary research. To support these goals, we've provided a wiki for the research group to use to facilitate communication of information privately, within the group. Wikis are gaining acceptance as tools for online collaboration across institutions (Butler, 2005). The OpenWetWare wiki (http://openwetware.org/), for example, was developed to support collaboration and information sharing in biology and biological engineering. To date, researchers in the group have used the wiki for share brief research descriptions, data collection plans, and posters prepared for project-wide meetings. In addition, Mann Library has used the wiki to share documentation and recommendations on formatting data and creating metadata.

We expect our local installation of Metacat to provide a means for the group to share both data and metadata privately within the group, prior to finalizing data sets and making them publicly available. It will also be the mechanism by which participants submit completed data and metadata to Mann Library, for publication to DSpace.

Early Lessons Learned

This collaboration is very much a work in progress, with researchers early in the process of data collection, and not yet ready to submit data and metadata for deposit in DSpace. Nevertheless, we've learned much from our work with this group that can inform our future efforts in this area.

One lesson we have learned is that in principle, some research communities are eager to share data both with their collaborators, and with the public, even though this is a new activity for many of them. This enthusiasm to share and archive data was manifested by an invitation to Mann Library to contribute to a separate grant proposal to the USDA for similar work on a larger scale including major watersheds of the northeastern United States and the Mississippi River watershed. More informally, several project participants have personally communicated their enthusiasm for this effort to staff at Mann Library.

We've also learned, with respect to the historic data described earlier, that even an apparently simple exercise in data formatting and documentation can take a significant amount of time and effort when undertaken long after the original data collection. Challenges to date have included migrating file formats (from Quattro Pro to Excel), assessing the impact of file format migration on data integrity, obtaining geographic coordinates for field sites visited 10-30 years ago, determining analytical methods and instrumentation used for chemical analyses, and resolving questions regarding apparent outliers in the data. Addressing these issues has required a good deal of personal communication between library staff and the data owner, research by the data owner to resolve questions of methodology and data quality, and cooperative work between library staff and the researcher to determine the exact location of historic field sites. A working knowledge of how ecological data are collected and of geographic information systems and maps has proven very useful in this effort. In addition, subject area knowledge has proven useful in spotting issues related to data quality. For example, in an intermediate version of one of the historic data sets, we noticed some water table elevation values that were outside the range of possible values for the area - an observation that might not have been made by someone without subject (and local) area knowledge. The data owner is working to reconstruct the correct values from field records. Subject area knowledge also made it possible for us to

catch and correct an error in documentation of units of measurement. Taken together, these observations make a strong case for early initiation of data documentation efforts by researchers.

From our preliminary knowledge of data collection efforts, we recognize that we will need to find appropriate procedures for dealing with simulation models. We will need to know what parts of a model should be included in a data submission package. For example, are input and output files from model runs necessary, to validate the performance of the model, or is the code alone sufficient? What are the best practices for preserving executable files? When scientists do "spreadsheet science", creating complex spreadsheets with embedded formulas and macros, what is the best way to document and preserve these files?

Finally, in practically all areas, we expect to find that "low" barriers to using new technologies may not be as low as we expect, as Lagoze et al. (2006) found in their experience with metadata creation for the National Science Digital Library. We've also found that some project participants have more easily (or enthusiastically) adopted the wiki as a medium for communication. Some users always upload or edit their own information, some request library staff to do this for them, and several report never using it at all. The public portal, while in theory editable by designated project participants, to date has only been updated by library staff, and this is likely to continue to be the case – an arrangement that is not sustainable or scalable over the long term.

Conclusions

Based on our early experience providing local (institutionally-based) support for research data curation, specific skills and strengths that are likely to prove useful for research libraries working in this area include: subject area knowledge, an appreciation of a discipline's cultural norms and history of sharing or curating data, an understanding of how research in the target discipline is conducted (from data collection to recording, transformation, analysis, and publication), a working familiarity with the most common software used in the discipline, existing standards and curation initiatives in the discipline, and the standard requirements of a discipline's most common funding agencies. On a more personal level, empathy and flexibility are very useful attributes.

We are optimistic about the potential role of research libraries. Likely future requirements that researchers submit a data management plan as part of a grant proposal mean that researchers will be looking for support to meet that requirement, and we are positioning ourselves to help provide it. In spite of what may prove to be a lower rate of adoption of some of the technologies in use in this collaboration, and some unanticipated extra work to assist with the documentation of the first data sets we received, we consider the investment of time and effort to be worthwhile, because many project participants view their experiences to date in this collaboration very positively and are likely to be willing cooperators in the future.

Acknowledgements

This work was funded in part by the National Science Foundation grant number 0437603 to Janet McCue and Barbara Lust. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Thanks to Jon Corson-Rikert, Janet McCue, and Anne Kenney for editorial comments.

References

ARL Workshop on New Collaborative Relationships. (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. Arlington, VA.

Butler, D. (2005). Joint efforts. *Nature, 438*(7068), 548-549.

Carlson, S. (2006). Lost in a sea of science data. *Chronicle of Higher Educ., 52*(42), A35-A37.

Caruso, B., Lowe, B.J., Corson-Rikert, J., & Devare, M. (2006). VIVO: Case study of an ontology-based web site. Arlington, VA. *Technical Report FS-06-06* 106-1.

Consultative Committee for Space Data Systems. (2002). Reference model for an open archival information system (OAIS). Washington, D.C.: CCSDS Secretariat.

Division of Ocean Sciences - National Science Foundation. (2003). Division of Ocean Sciences data and sample policy No. NSF 04-004.

Fegraus, E.H., Andelman, S., Jones, M.B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America, 86*(3), 158.

FGDC Biological Data Working Group. (1999). Content standard for digital geospatial metadata - biological data profile, FGDC-STD-001.1-1999. Washington, DC: Federal Geographic Data Committee. Online: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf. Retrieved March 12, 2007.

Glover, D.M., Chandler, C.L., Doney, S.C., Buesseler, K.O., Heimerdinger, G., & Bishop, J.K.B., et al. (2006). The US JGOFS data management experience. *Deep-Sea Research Part II: Topical Studies in Oceanography, 53*(5-7), 793-802.

Green, A., & Gutmann, M.P. Building partnerships among social science researchers, institution-based repositories, and domain specific data archives. Oline: http://hdl.handle.net/2027.42/41214. Retrieved March 12, 2007.

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing, 5*(5), 59-68.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work: CSCW: An International Journal, 15*(4), 321-358.

Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., & Saylor, J. (2006). Metadata aggregation and automated digital libraries: A retrospective on the NSDL experience. *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries,* Chapel Hill, NC, USA. 230-239. Online: http://doi.acm.org/10.1145/1141753.1141804. Retrieved March 12, 2007.

Lord, P., & Macdonald, A. (2003). *e-science curation report data curation for e-science in the UK: An audit to establish requirements for future curation and provision.* JISC Committee for the Support of Research. Online: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf. Retrieved March 12, 2007.

Messerschmitt, D. G. (2003). Opportunities for research libraries in the NSF cyberinfrastructure program. *ARL,* (229), 1-7.

RLG-NARA Task Force on Digital Repository Certification. (2005). In n/a (Ed.), *An audit checklist for the certification of trusted digital repositories* (1st ed.). Mountain View, CA: Research Libraries Group (RLG). Online: http://www.rlg.org/en/rlgnara-repositoriescheck-list.pdf. Retrieved March 12, 2007.

Strategic plan for the USGS national biological information infrastructure (NBII)(2005). Online: http://www.nbii.gov/about/pubs/NBII%20Strategic%20Plan.pdf. Retrieved March 12, 2007.