Integrating Image-Based Research Datasets into an Existing Digital Repository Infrastructure

Hannah Tarver

Digital Projects Unit, University of North Texas, Denton, USA

hannah.tarver@unt.edu

Mark Phillips

Digital Libraries Division, University of North Texas, Denton, USA

mark.phillips@unt.edu

Integrating Image-Based Research Datasets into Existing Digital Repository Infrastructure

In 2011, the University of North Texas (UNT) Libraries partnered with researchers in the university's academic departments to describe and provide access to items not traditionally included in the UNT Libraries' systems. Including more than 1,400 items apiece, the two projects are considered active datasets by their respective users. Each collection provided new challenges in harmonizing partner, metadata, and end-user requirements. This article discusses the projects, workflow for defining requirements, and final implementation in the UNT Digital Library. These collections serve as a model for integrating other research projects easily and inexpensively into a repository infrastructure.

Keywords: institutional repositories, digital libraries, grant projects, digital images, data management

Introduction

The University of North Texas (UNT) Libraries Digital Projects Unit (DPU) works with many departments and organizations on the university campus to digitize and make available items created at or owned by the university by putting them into the UNT Digital Library. Items in the Digital Library range from government documents and scholarly publications to videos and images used for instruction and currently total more than 60,000 objects available for use.

The UNT Digital Library has steadily added new collections and content types since its creation in the mid-2000s, primarily focusing on printed materials or items held by the UNT Libraries. The DPU was eager to extend past these traditional practices to work more directly with faculty who are creating and acquiring digital content for their research, since this would dovetail nicely with other initiatives at UNT, such as an Open Access Policy and the creation of the UNT Scholarly Works repository. Since faculty research data can take a variety of forms in a number of formats, the UNT Libraries feel

that managing these data for the university is a priority, and ways to accomplish that are under investigation.

During 2011, the Digital Projects Unit partnered with two UNT professors working on grant projects; in both cases, the professors had obtained grant funding for projects that resulted in image data. The Digital Library provided an existing framework to house the digital files and make them available to the research community and the public-at-large.

The first project was conducted in partnership with the Elm Fork Natural Heritage Museum, which is housed on the UNT campus and functions in connection with the environmental and biological science departments. The museum received a grant from the Institute of Museum and Library Services (IMLS) to digitize a large collection of specimens for species of mussels found in Texas. The goal was to make high-quality images publicly available to help researchers study mussel species without having to handle fragile original specimens or collect their own specimens, particularly since several of the species are currently or may soon be on the endangered list. In this case, the grant was written with the intention of placing the final images in the UNT Digital Library.

The second project involved a partnership with a research team in the UNT School of Merchandising and Hospitality Management. The group received a grant from the United States Department of Agriculture (USDA) to study the foods that middle school students choose to put on their trays at lunch and what they actually eat, in a project titled "Testing a Food Choice Innovation for Middle School Cafeterias." To collect information for study, the researchers labelled cafeteria trays with unique numbers and created a station where students could photograph their trays before and after eating lunch in order to document plate waste; the resulting research data includes

nearly 3,000 images of school lunch trays. For this project, the grant proposal did not specify how the images would be managed; the researchers were referred to the Digital Library as a possible host after the images were captured.

Both of these collections represent significant new partnerships between the digital repository managed by the UNT Libraries and research professors on the university campus engaged in grant-funded projects.

Literature Review

Although the role of academic libraries has always been to curate information and provide access to resources that support the work of parent institutions, this role is broadening to incorporate digital resources and to accommodate new data requirements. As Shuler noted, digital information has become integrated into library values, and relying on physical, print resources is no longer a responsible way for libraries to provide information access.[1] Many authors have noted this changing role of libraries including the need to provide portals or pathways to digital resources,[2] the ways that libraries can investigate the use of digital tools[3] and the commitment that libraries must have to integrating these services.[4] Additionally, authors have noted the large variety of item types that institutional repositories house or expect to house including the survey by Lynch and Lippincott of early U.S. repositories;[5] Henty's analysis that academic libraries use repositories to store images, sound files, and other data;[6] and the discussion by Hulse, Cheverie, and Dygert includes the difficulties in creating a repository for the Washington Research Library Consortium when some members wanted to contribute media files in addition to text.[7] Some authors also discussed research possibilities such as creating "shariums" within the university community[8] to promote collaboration across disciplines and with peers located internationally.

In fact, ways that libraries can actively involve themselves in the research process are being addressed on many fronts; some institutions, including the Massachusetts Institute of Technology (MIT), have even initiated repositories entirely or in part to house and support in-house research.[9]

Henty discussed the challenge for libraries to become integral parts of "eResearch" as data stewards based on surveys of institutions in Australia.[10] This interest has also been addressed by Queensland University of Technology as library staff members look for more specific ways to incorporate eResearch into their services.[11] In the United States, the Association of Research Libraries (ARL) has been exploring this question for several years, creating task forces and sponsoring workshops to consider the issues involved. In 2006, ARL held a workshop to consider "the role of academic libraries in the digital data universe"; the findings state the need for academic libraries to be directly involved in storage, preservation, and curation of research data, particularly in relation to science and engineering activities.[12] The ARL Joint Task Force on Library Support for E-Science noted in its 2007 report that librarians already have skills and tools related to metadata creation and item storage that could serve as a foundation for supporting the curation of digital research data.[13] Henty also noted that the role of academic libraries in research need not be limited exclusively to science-related endeavors.[14]

As one step toward seeing where academic libraries may play a role in research, several institutions and committees have studied ways that researchers and faculty manage and archive their data independently, including The University of Queensland, The University of Melbourne, and Queensland University of Technology in Australia;[15] the University of Houston in Texas;[16] the University of Colorado Boulder;[17] and the Joint Information Systems Committee in the United Kingdom.[18] The research in

Australia showed that at some institutions, many of the primary investigators (PIs) involved in grant research did not have data management plans in place.[19] Library researchers at the University of Houston discovered that although many PIs claimed to have responsibility for data management in their research, it was apparent that most of them did not fully understand what was meant by "data management."[20]

Despite the lack of understanding, many funding agencies, particularly those associated with the U.S. government are increasingly expecting data management plans and access to research data or findings on a long-term basis as a grant criterion.[21] For example, the United States National Science Foundation (NSF) is responsible for roughly 20% of basic federal research funding awarded to academic institutions.[22] The NSF requires a data management plan for all grant applications and part of the documentation must show that investigators will share "the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."[23] The U.S. National Institute of Health (NIH) has a similar policy, stating, "To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible."[24]

Data management also includes the creation of useful metadata to describe research and make data findable. The National Information Standards Organization (NISO) documentation regarding what makes digital collections "good" includes metadata principles regarding curation, preservation, archivability, and persistence.[25] Making information findable can be more complicated. The third metadata principle in the NISO documentation, regarding content standards, mentions that metadata should use authority control and vocabularies in line with the users of the digital objects,[26] but

in a digital environment, the expected user-group may not be the exclusive or majority of users who actually access and utilize the items. Even if a single user-group becomes the focus of the metadata, institutional repositories may have items crossing several specialized disciplines; a study of the repository at John Hopkins University noted that this could be a particular problem since the cataloguing staff were practiced at making useful metadata but did not have the level of domain knowledge that contributors wanted.[27] After a study of metadata at various digital libraries worldwide, some researchers noted that digital collections that continue to acquire new items and develop additional collaborations will have to revisit established metadata guidelines as new issues appear.[28] All of these factors in combination with the skill sets and mission of academic libraries point toward more involvement between university grant researchers and institutional repositories or curation of data in library-held digital resources.

Establishing a Workflow

Establishing a workflow for each of the new collections involved project meetings to outline imaging standards and file-naming conventions. This ensured that the digital images met the UNT Digital Library standards and were appropriately organized. Because both projects were sourced from digital cameras, DPU staff instructed researchers at the start of the projects on how to provide the highest quality images. The mussel project resulted in high-resolution tiff images while the lunch tray project produced lower-resolution jpeg images.

One important step in discussing these projects with the grant researchers was explaining that, as the experts, they needed to organize the files into logical groupings to become individual digital "objects." Although the lunch trays were reasonably straightforward – each digital object has one before and one after picture for a total of two images – some of the mussel specimens had as few as two and as many as six

images for a single digital object. In both cases, it was up to the research staff members to sort through the digital files and ensure that the appropriate files had been matched together.

The DPU team has a standard practice of using collection-relevant file-naming conventions for all projects that will be moved into the Digital Library. This proved useful with both projects as each group had unique identifiers for distinguishing individual items. Managing digital items for this process involves the use of a unique identifier for each digital object (e.g., a call number, accession number, or other assigned number) with an appended two-digit zero-padded number for each file name to denote the sequence of images. This ensures that the lunch trays always display before then after, and that the mussel specimen perspectives are always displayed in the same order. The Elm Fork researchers chose to use a catalog number that was already assigned to the mussel specimens in a separate database (see figure 1); the hospitality researchers constructed identifiers that included a code for the school, the date, and individual tray for each identifier (see figure 2).

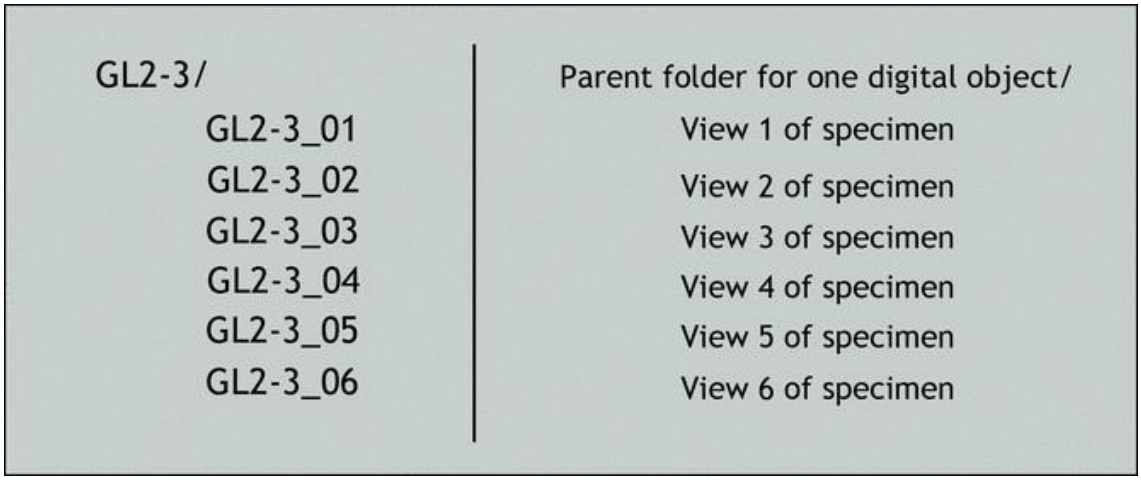


Figure 1. Example of file structure for mussel specimens.

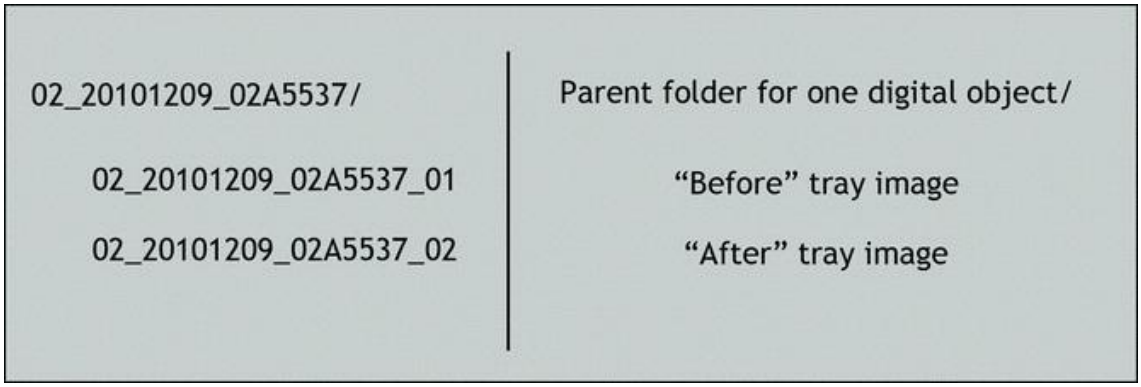


Figure 2. Example of file structure for lunch tray images.

Throughout both projects the DPU staff answered questions and moved example items into a staging repository so that faculty members could see how their imaging and organization decisions would affect the end product.

Metadata Creation

When project "data" consists of images rather than numbers or text, metadata is invaluable in quantifying the results; for both of these projects, metadata was also a key component for ensuring that items could be easily discovered and used, both by the public and by the grant researchers. However, there were special challenges in meeting the metadata requirements for each scenario.

Partner Requirements

In each of the projects, the researchers managing the digital objects expressed specific information needs for organizing the collections online. One of the main concerns for the grant holders in the Elm Fork Mussel project was that the primary users for the collection of specimen images would be other specialists who would have exacting standards and searching needs. This meant that they wanted to use highly-technical language and scientific specifications for describing each of the specimens. However,

since much of the traffic to the Digital Library comes from global users searching for keyword terms and landing on item pages, the specialists' requirements had to be balanced with text that would help orient non-specialists and that would roughly match the style already established for describing items within the digital collections.

Although the project documenting middle school plate waste did not require the same level of inherent technical expertise to understand and use the images, the project did have other special requirements. First and foremost, the researchers were concerned about maintaining the anonymity of the participants who had been involved in the study. To accommodate this, the metadata for this project included generic, nearly-identical records for every image-pair in the collection with information that was as broad as possible. For example, as the coverage location, the records list "United States – Texas" but without specifying the city or even the county. The grant researchers also needed a way to quantify findings from the images to create numerical data for analysis, although they did not have specific expectations about how to accomplish that goal using metadata information.

Metadata Requirements

After the researchers expressed their needs for collection use, DPU staff considered the best ways to meet as many requirements as possible while working within the metadata guidelines established for the Digital Library. One benefit of adding items to larger digital holdings is that metadata specialists know the system requirements and capabilities, techniques for handling information organization across diverse collections, and ways to improve the quality and functionality of the information that researchers already have.

For the lunch tray project, the researchers involved in the grant had an opportunity to leverage features that are already available in the structure of the Digital Library as a means of using and analyzing their visual data in a quantifiable way. Many of the fields in the Digital Library are automatically linked (such as series title and keyword), so information was added in some of those fields as a way of easily counting or narrowing the collection to a subset for analysis. Since the project was conducted at two separate schools, there is a series title for each of the schools (School #1 and School #2, rather than identifying names); this makes it possible to easily isolate all of the data related to one school or the other.

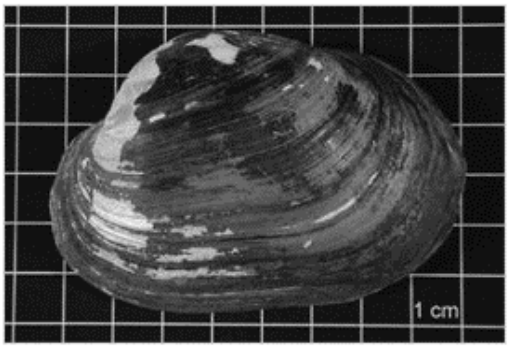
Currently, the grant researchers are independently developing a controlled vocabulary to code images by adding the terms as keywords to the records. For example, they may use the term "canned fruit innovation" to label the subset of images that were captured when canned fruit was introduced to the available foods in the cafeteria. The researchers will add appropriate keywords that describe both the foods that were chosen and the variables that were introduced at certain stages; in the Digital Library, those terms will automatically be linked as a term search which will give a quick count of relevant records (providing quantifiable data) and a way to narrow the dataset to a specific group of images to review.

End-User Requirements

Ultimately the most important concern for both the content holders and the digital hosts is providing access to items. Since digital items have a broad user base, DPU staff try to balance the metadata and system requirements with more specific information that researchers need in relation to the items. It can be easy for content holders to become focused on the ways that they would like to organize and manage their own discreet

collections without having the same perspective that metadata specialists – or traditional catalogers – have when organizing multiple collections and dozens of item types in the same system.

For example, when formulating records in the Elm Fork mussel collection, the metadata needed to address the need for balance between experts and laymen. In this case, the researchers and DPU staff cooperatively formulated a standardized content description statement written in sentences similar to other record descriptions, but that included all of the information that the specialists deemed most necessary. The same wording was used for every description to allow users to search for specific criteria by entering exact phrases. The description contains all relevant information about the specimen and includes the number of valves, shell shape and thickness, internal and external colors, beak and external sculpturing, shell length, shell condition, and location of collection (see figure 3). The precise wording makes it easier for experts to find specific specimens based on familiar criteria; the sentence form helps non-experts to understand what they are viewing.



Description: One preserved mussel specimen including both the left and right valves. The specimen exhibits an oval shape; moderately thick shell; double looped beak sculpturing; white internal coloring; brown external coloring; no external sculpturing. Collected in the Trinity basin. The specimen measures between 60 - 100 mm in length and was assessed to be relatively-recently dead when collected.

Figure 3. Example content description for a mussel specimen.

Collaborative Metadata

Perhaps the most challenging part of projects similar to these is finding a way to create useful metadata records when the metadata specialists are not subject experts. In the

case of the Elm Fork project, in particular, the information available to DPU staff for the collection was highly specialized (e.g., beak and valve descriptions for mussels), so trying to understand how best to translate the values for metadata records required some trial-and-error discussions with the researchers. It was also important in that case because the grant team intended for the primary audience to be other subject experts and did not initially see the value in making the specimen records more understandable for public users.

Another communication issue in determining metadata for the mussels was that the researchers were most accustomed to cataloguing specimens as database entries. An important turning point in the metadata discussion occurred when the researchers understood that, for a repository, the metadata needed to look more like a field guide and less like a database entry, because the records would not be acting as surrogates for the physical objects or functioning in a database list.

Metadata specialists who understand formatting guidelines can also improve the end-user experience since many of the metadata fields in the UNT Digital Library (including series title and subjects) drive the browsing functionality. In some cases, such as the lunch tray project, DPU staff were able to make suggestions for metadata records to improve the usability of the collection by adding information in ways that researchers had not considered. However, expressing the importance of controlling metadata can be a hurdle to collaborating with researchers when an established metadata schema might not make the fine distinctions that they would prefer. The challenge is shifting the focus of researchers from describing a single collection toward larger implications, such as building conceptual connections across unrelated collections and finding compromises. In each of these projects, the collaborative

metadata formulated by subject experts in conjunction with DPU staff ended up being more useful and robust than if it had been developed by any one group alone.

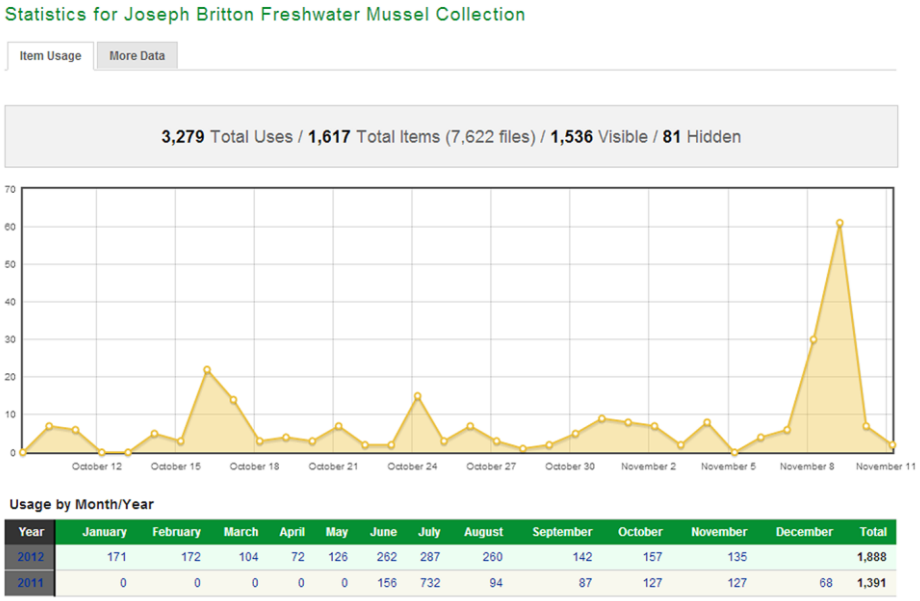
Implementation

In each project a significant period of time separated the first contact initiated by the faculty members in their respective grant projects and the point when the Digital Projects Unit saw the final product of the project. Although DPU staff assisted with technical requirements, staff members were not involved in generating or collecting the image data and only re-entered the project workflow when all of the images were organized and ready to go into the repository.

Each project used external Universal Serial Bus (USB) hard-drives to deliver files from researchers to the Libraries. Once the files were transferred to the Libraries they were loaded onto network storage volumes used for staging digital projects and collections. DPU staff executed a number of quality control procedures to ensure consistency of the files as early as possible. Simple BASH (Bourne Again SHell) scripts were used to check the file names used and to verify that they matched the pre-defined file naming conventions. Then, staff employed the image manipulation suite ImageMagick to check that each image file matched the pre-defined imaging specifications.

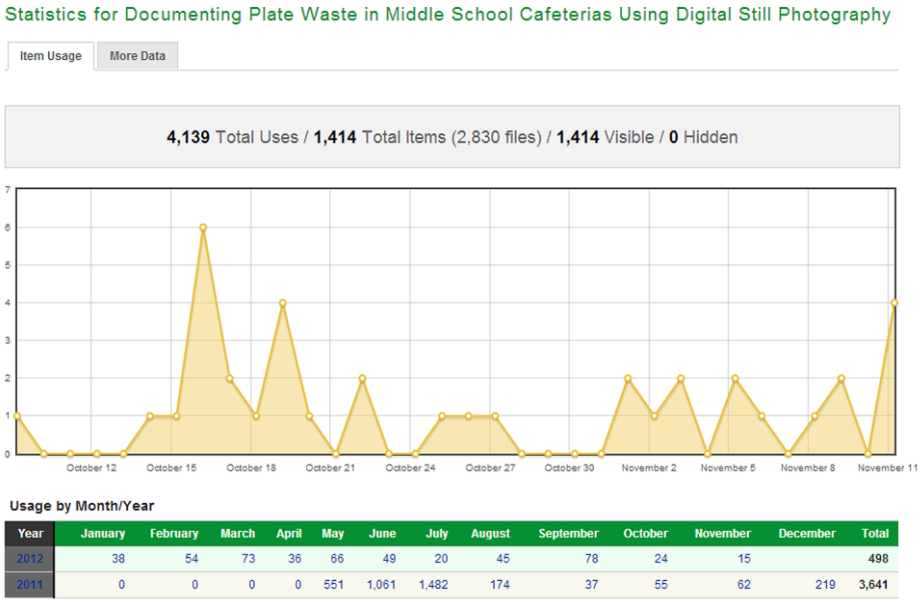
Once these quality control steps were complete and any anomalies and errors were corrected, the items were staged for ingest into the repository. The Digital Library has a standard ingest workflow which creates submission packages, extracts preservation and structural metadata, and finally creates Web derivatives for delivery to the end user. These standard workflows allowed all content to be processed for both collections with a minimal amount of staff time from the DPU. The items are now

online and a variety of users access both the specimens (see figure 4) and the lunch tray images (see figure 5) dozens of times per month.



Year	January	February	March	April	May	June	July	August	September	October	November	December	Total
2012	171	172	104	72	126	262	287	260	142	157	135		1,888
2011	0	0	0	0	0	156	732	94	87	127	127	68	1,391

Figure 4. Statistics for the mussel specimen digital collection through mid-November 2012.



Year	January	February	March	April	May	June	July	August	September	October	November	December	Total
2012	38	54	73	36	66	49	20	45	78	24	15		498
2011	0	0	0	0	551	1,061	1,482	174	37	55	62	219	3,641

Figure 5. Statistics for lunch tray image digital collection through mid-November 2012.

Conclusion

Working on these two grant projects has been mutually beneficial for the Digital Library and the researchers that partnered with the Digital Projects Unit. Adding these types of items helps to build and diversify the Digital Library and to start the groundwork for a broader connection with research initiatives at the university. Meanwhile, both researchers participating in the grants have a sustainable and accessible repository to store their research data, giving them better ways to quantify and analyze their results, share their work with others, and meet data management expectations from granting agencies without the need for personal or departmental resources.

Lessons Learned

Each grant project has special requirements in terms of data management plans required by granting agencies, the kinds of files generated as data, and the metadata needs that researchers may have in order to both promote and protect their work. It is important for repository managers to address these needs at the start of the project and maintain clear communication with researchers to ensure that the end product works for all parties. Similarly, it can be helpful to consider the terminology and usual methods for storing data in specialty disciplines. For example, the Elm Fork Mussel researchers primarily use specimen databases to store and find specimens; the discussion about metadata requirements and finding reasonable compromises regarding their project proceeded more smoothly once there was a general understanding that the Digital Library does not function as a specimen database, in terms of finding and using information.

Future Implications

Since many grant-funding agencies require researchers to implement a data management plan for their findings, there is an opportunity for academic libraries to support that need and, at the same time, to curate digital collections of research data. Librarianship focuses on matching users with resources and information. Researchers often have the same goal, though they may have narrower ideas about "who" comprises their user groups. The skills already held by catalogers and metadata specialists to organize information for diverse items and make them findable means that academic libraries are perfectly situated to collect digital objects and improve their functionality for internal and global use. Many academic libraries also have some digital framework already in place. These two projects demonstrate how academic libraries may be able to offer support for some research grants by integrating data into existing institutional repositories without expending time and resources to build or maintain new systems and specialized software.

Suggestions for Further Research

Although these projects demonstrate ways that academic libraries may be able to accommodate grant research data in the form of images in existing repositories, many grants generate data other than images. More research could be conducted on ways that repositories can manage diverse items and increase their flexibility. Additionally, since many universities focus heavily on participating in scholarly research and securing grants, it may be beneficial to compile more general studies on how university libraries can assist not only as resources for creating data management plans, but as key data repositories for their parent institutions.

Notes

1. John Shuler, "Academic Libraries and the Global Information Society," *The Journal of Academic*

Librarianship 33, no. 6 (December 2007): 710, doi:10.1016/j.acalib.2007.09.018.

2. John Akeroyd, "The Future of Academic Libraries," *Aslib Proceedings* 53, no. 3 (March 2001): 82.
3. Clifford Lynch, "Where Do We Go From Here? : The Next Decade for Digital Libraries," *D- Lib Magazine* 11, no. 7/8 (July/August 2005), http://www.dlib.org/dlib/july05/lynch/07lynch.html (accessed February 21, 2012)
4. Rachel Heery and Andy Powell, *Digital Repositories Roadmap: Looking Forward* (Bath, UK: OPUS: University of Bath Online Publication Store, 2006), 8. http://www.jisc.ac.uk/media/documents/programmes/reppres/reproadmap.pdf.
5. Clifford Lynch and Joan K. Lippincott, "Institutional Repository Deployment in the United States as of Early 2005," *D-Lib Magazine* 11, no. 9 (September 2005), http://www.dlib.org/dlib/september05/lynch/09lynch.html.
6. Margaret Henty, "Dreaming of Data: the Library's Role in Supporting E-Research and Data Management." Presented at the Australian Library and Information Association Biennial Conference, Alice Springs, NT Australia, September 2-5, 2008, 2, http://apsr.anu.edu.au/presentations/henty_alia_08.pdf.
7. Bruce Hulse, Joan F. Cheverie, and Claire T. Dygert, "ALADIN Research Commons: a Consortial Institutional Repository," *OCLC Systems & Services: International Digital Library Perspectives* 23, no. 2 (2007): 167-168, doi: 10.1108/10650750710748469.
8. Jeffrey Pomerantz and Gary Marchionini, "The Digital Library as Place," *Journal of Documentation* 63, no. 4 (2007): 523, doi:10.1108/00220410710758995.
9. Patsy Baudoin and Margret Branschofsky, "Implementing an Institutional Repository: The DSpace Experience at MIT," *Science & Technology Libraries* 24, no. 1/2 (June 2004): 32, uri: http://hdl.handle.net/1721.1/26699.
10. Henty, "Dreaming of Data," 4.
11. Jennifer Thomas, "Future-Proofing: the Academic Library's Role in E-Research Support," *Library Management* 32, no. 1/2 (2011): 38, doi:10.1108/01435121111102566.
12. Association of Research Libraries, *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, September 26-27, 2006, 42, http://www.arl.org/bm~doc/digdatarpt.pdf.

13. Association of Research Libraries, Joint Task Force on Library Support for E-Science, *Agenda for Developing E-Science in Research Libraries,* (2007), 13, http://www.arl.org/bm~doc/ARL_EScience_final.pdf.
14. Henty, "Dreaming of Data," 3.
15. Ibid, 5.
16. Christie Peters and Anita Riley Dryden, "Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston," *Science & Technology Libraries* 30, no. 4 (2011): 387-403, doi:10.1080/0194262X.2011.626340.
17. Kathryn Lage, Barbara Losoff, and Jack Maness, "Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder," *Libraries and the Academy* 11, no. 4 (2011): 915-937, http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.4.lage.html.
18. Heery and Powell, *Digital Repositories Roadmap*, 16.
19. Henty, "Dreaming of Data," 6.
20. Peters and Dryden, "Assessing the Academic Library's Role," 394.
21. Lage, Losoff, and Maness, "Receptivity to Library Involvement in Scientific Data Curation," 916.
22. "US NSF – About Awards," last modified February 29, 2012, accessed April 18, 2012, http://www.nsf.gov/awards/about.jsp
23. "US NSF – Award and Administration Guide: Chapter VI – Other Post Award Requirements and Considerations," last modified November 8, 2011, accessed April 18, 2012, http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4
24. "NIH Data Sharing Policy and Implementation Guidance," last modified March 5, 2003, accessed April 18, 2012, http://grants.nih.gov/grants/policy/data_sharing_guidance.htm
25. National Information Standards Organization, "Metadata" in *A Framework of Guidance for Building Good Digital Collections*, 3rd ed. (Baltimore, MD: National Information Standards Organization, 2007), http://framework.niso.org/node/24 (accessed April 23, 2012).
26. National Information Standards Organization, "Metadata Principle 3" in *A Framework of Guidance for Building Good Digital Collections*, 3rd ed. (Baltimore, MD: National Information Standards Organization, 2007), http://framework.niso.org/node/246
27. John W. Chapman, David Reynolds, and Sarah A. Shreeves, "Repository Metadata: Approaches and

Challenges," *Cataloging & Classification Quarterly* 47, no. 3-4 (2009): 316, http://dx.doi.org/10.1080/01639370902735020

28. Marcia Lei Zeng, Jaesun Lee, and Allene F. Hayes, "Metadata Decisions for Digital Libraries: A Survey Report," *Journal of Library Metadata* 9, no. 3-4 (2009): 175, http://dx.doi.org/10.1080/19386380903405074