A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot

Lisa Johnston, Research Data Management and
Curation Lead, University Libraries
University of Minnesota - Twin Cities

Published January 19, 2014 at http://hdl.handle.net/11299/162338

Cite as: Johnston, L.. (2014). A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot. University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/162338

Preface

This internal report details the results of the Data Curation Pilot run at the University of Minnesota (UMN) by the University Libraries from May-December 2013. The author carried out this project as a cohort member of the 2013 President's Excellence in Leadership (PEL) program, thanks to the generous support of sponsors John Butler and Karen Williams, Associate University Librarians. In 2013, the PEL program refocused its project aspect to be department-centric, rather than the previous team-led approach to a common university problem. Therefore, units were invited to select one participant for the 2013 PEL cohort to work on a project to directly benefit their unit's goals and mission. The project proposal was delivered to the sponsors and approved by University Libraries Cabinet on May 21, 2013 (see full proposal in Appendix A). In addition to unit-based sponsors, the PEL program paired "PEL Circles" of 4-5 PEL participants with a PEL Mentor that met monthly throughout the program. The author's mentor was Dr. Brian Herman, UMN's Vice President for Research. A presentation on the data curation pilot was delivered to Dr. Herman and the PEL circle on August 22, 2013. The results of the project were delivered at the December 3, 2013 PEL closing reception that included leaders from across the university.

Executive Summary

The 2013 Data Curation Project set out to test and expand the University Libraries' programmatic and technical capacities to support research data management needs on campus by establishing a fixed-term data curation pilot. This pilot utilized our current suite of services and expertise in the University with the objective of developing "workflows" for curating a variety of types of research data. Specifically, in eight months, this project resulted in 1) a data curation workflow utilizing existing university resources; 2) five pilot research datasets that were solicited, selected, and curated for discovery and reuse in the libraries' digital repository, the University Digital Conservancy; and 3) and a summary report describing the successes and shortcomings of this approach.

The University of Minnesota's 2013 Data Curation Pilot's primary task was to develop and implement curation workflows for 3-5 examples of research data. A call for proposals to participate in the pilot went out in the summer of 2013 and was open to researchers on campus whose data met a variety of criteria (including openness to the public). In response to the call, 16 proposals were received, and five were selected to represent a variety of disciplines and data types. These were:

- Engineering Data: GIS data from reverse-engineering print transportation maps created by David Levinson (Civil Engineering, CSE).
- Health Sciences Data: Excel and .csv data from periodontal clinical trials created by James Hodges (School of Public Health, ACH).
- Interdisciplinary Data: Excel data of chemical traces found in Minnesota lakes created by Bill Arnold (Civil Engineering, CSE).
- Natural Resources Data: SPSS data from online tourism surveys created by Lisa Qian (Forest Resources, CFANS and Extension).
- Social Sciences/Humanities Data: Video and transcription files from Ojibwe conversations created by Mary Hermes (Curriculum and Instruction, CEHD).

Next, a detailed treatment process was developed for each of the pilot datasets and formed the bases of the overall curation workflow. To accomplish this, the pilot involved the expertise of archival, digital preservation, and metadata and cataloging staff in the library, as well as data experts from the university, to curate the digital research data while utilizing existing tools, such as the institutional repository. The project resulted in all five of the pilot dataset being successfully curated for discovery and reuse in the University's institutional repository, the University Digital Conservancy, at the persistent URL, http://purl.umn.edu/160292. To supplement this process, pre- and post-curation interviews took place with the participating data authors in order to determine the extent of their perceived need

for data curation services and the resulting success or shortcomings of the final curated product.

This report summarizes the steps taken to curate the datasets in the pilot. In addition to the specific dataset treatments, an overall data curation workflow is presented that outlines the steps needed for any dataset. A discussion of this process provides some useful lessons learned. For example:

- Through the interview process, it became evident that several faculty were less concerned with archiving their data for others to access, or even meeting federal mandates, than with finding a permanent home for their data to "live on" with restricted access.
- As future services are developed, it will be important to consider the variety of software, and expertise to use the software, required for data curation. Important software for this study included statistical tools (SPSS, R) and GIS software (ArcGIS).
- The researchers interviewed did not have ready documentation to provide with their datasets. In several cases, readme.txt files were written by curation staff to supplement the data.
- All of the datasets included some aspect of ownership and intellectual property considerations, even though the pilot was explicit that all submissions were dataset ready for public consumption and reuse.

As a result of this project, the University Libraries now hold a more realistic sense of the overall capacities and expertise needed to develop a sustainable data curation service model. Additionally, the Libraries are better prepared to fine-tune and implement selected recommendations from previous assessments and committee reports. Due to variables of scale, domain-specific data requirements, and diversity of domain culture and practices, the success of such a model will likely depend upon strong collaboration among interdependent service providers on campus. To be successful, significant capacities in areas data management and curation, infrastructure, and domain knowledge must coalesce in operationally effective ways that minimize barriers to and demands on researchers.

Introduction

Over the last several years, researchers and administrators at the University have developed a growing awareness of and desire for long-term access to digital research data. One recent driver is the February 22, 2013 memorandum by the White House (OSTP)[1] asking federal agencies to develop policies "requiring researchers to better account for and manage the digital data resulting from federally funded scientific research." Several federal funding agencies[2] already require investigators to include a plan for how they will share research data, but this new step mandates that resulting data is "publicly accessible to search, retrieve, and analyze." The implications of this policy, such as the need for federal support versus institutional support, are still unfolding. Initiatives within the academic library community (e.g., the SHARE initiative, http://www.arl.org/share) anticipate and are positioned to respond to an increased campus needs for data management and repository services.

To expand our programmatic efforts while taking into account a variety of needs[3] from scholars across the disciplines, the UMN Libraries initiated a Data Curation Pilot in the spring of 2013. This pilot aimed to test existing institutional capacities in support of digital curation, including appraisal, ingest, arrangement and description, metadata creation, format transformation, dissemination and access, archiving, and preservation. Data curation also enables UMN researchers to comply with pending government requirements to make the digital data associated with federal grants available for sharing and reuse. These federal requirements will affect 68% of the grants (e.g., NIH, NSF) at the University of Minnesota, according to 2012[4] data.

This pilot utilized our current suite of services and expertise in the Libraries by documenting our curation steps in the form of "curation workflows" for a variety of types of research data. Additionally, the project incorporated change management techniques to engage campus stakeholders in data curation through dialogue and events. As a result, the pilot developed a set of potential treatment processes and workflows for example datasets that the Libraries and its institutional partners might extrapolate towards a full-fledged service model for data curation.

[1] More at http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-

[2] See https://www.lib.umn.edu/datamanagement/funding for a list.

[3] Finding included in the unpublished 2012 report, "Deepening Support for Research: A Strategic Agenda for E-Science in the UMN Libraries" and the 2012 *Journal of Library Administration* article "Developing E-science and Research Services and Support at the University of Minnesota Health Sciences Libraries" available at http://purl.umn.edu/159983.

[4] Based on fiscal year 2012 data at http://www.research.umn.edu/news/stats.html#.Uo1F6WSG07t

Objectives of the Pilot

The main goal of the project was to identify, select, and pilot data curation services for 3-5 research data examples. Objectives included:

- Explore and document faculty expectations and needs through interviews and engagement activities.
- Involve University staff (digital curators, digital technologists, data management staff, and subject-specific librarians) to establish best practices for data curation treatments and to develop a curation workflow based on the example datasets.
- Curate the 3-5 pilot projects using existing infrastructure (e.g., the University Digital Conservancy).

In addition, the project set out to document the data curation workflow and to assess this result, including the successes and the shortcomings of the approach.

As a result of this project, the University Libraries are better prepared to make recommendations and implement sustainable data curation services in the future. Also, with change management and engagement techniques to address both internal data curation for library staff and external data curation for campus, this pilot demonstrates the role that the library can play in supporting research data across the University.

Literature Review

This literature review focuses on three primary issues involved with the 2013 Data Curation Pilot:

1. The concept of data curation is still an emerging topic within library science, archives, and information sciences. Few academic libraries are successfully offering data curation services, according to a recent ACRL white paper (Tenopir, Britch, and Allard, 2012). This review will highlight several exemplary models of data curation that can be grouped into two categories: models that incorporate the data and/or research life cycle and models that incorporate digital object curation workflows.
2. The review will open up the topic of curation to encompass archival best practices for all digital objects, not just data. The archival community has dealt with curation issues in the print and analog for centuries, and those lessons learned translate well into the digital realm.
3. Finally, the review will discuss data curation practices and workflows in disciplinary and institutional data repositories, including the implementation of pilots like ours.

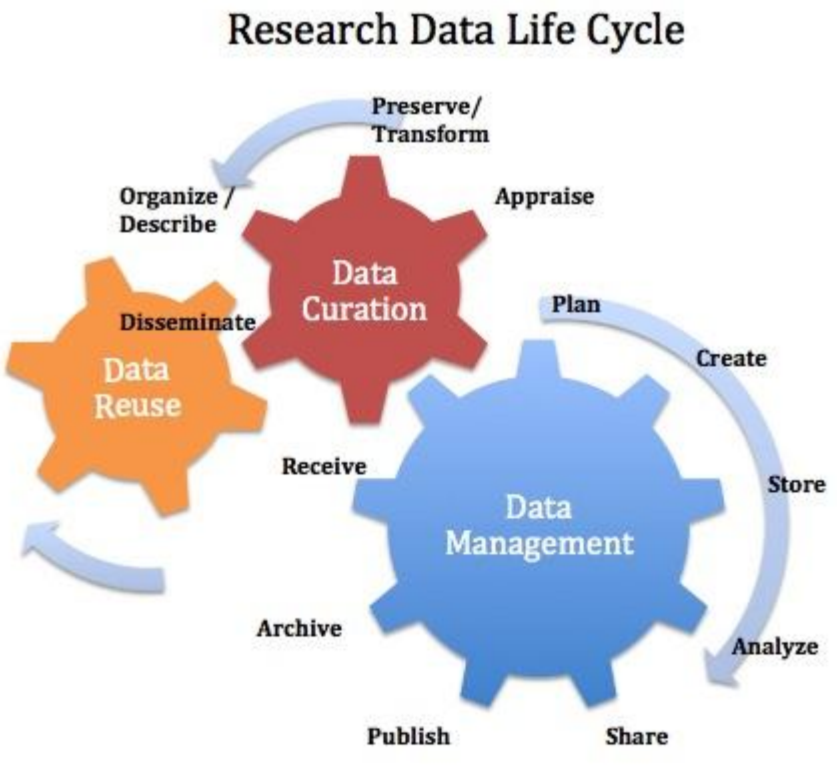
This is by no means an exhaustive review, but a selection of examples to illustrate the work being done by many practitioners and theorists who are tackling data curation today.

Data Curation and the Models That Illustrate It

Data curation is described by the University of Illinois's iSchool Data Curation Specialization as "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation" (UIUC, 2013).

However, for the purposes of this pilot project, it is useful to visualize the processes of data curation as part of the larger *research data life cycle* (see Figure 1). For example, the steps of creating, analyzing, storing, and publishing the results of data primarily happen much earlier in the research life cycle, before the curation phase. Additionally, the steps involved in data curation may be done by a third party (the repository curators) rather than the principal authors of the data. Therefore, if a data curation service is to be successful, the full data life cycle must be taken into account, and collaboration between the curators and the authors should begin as early as possible in the research life cycle.

Figure 1: Conceptual Model of the Research Data Life Cycle



Using life cycle models to visualize data services in academic libraries is becoming the norm (Carlson, 2014). These models take into account the full research life cycle and give primacy to the data management aspects. The models, of course, vary. Excellent examples can be found represented in several forms, such as a linear process (DDI Alliance, 2004; Humphrey, 2006; University of Virginia Libraries, 2013), a circular process (ICPSR, 2012, p. 8; DataOne, 2013), or a combination of the two (UCF, 2012; Rohrs, 2013). The last two examples from the University of Central Florida Libraries and New York University are notable as they not only include the stages of the research life cycle, but also include the wide variety of services and service providers (e.g., the libraries, information technology, grants administration, etc.) that researchers might encounter in their academic research settings.

Models can also include more detailed steps, which help to illustrate the complexity of the data curation process for researchers and others less familiar with it. For example, the workflows model described in the UK Data Archive report (Van den Eynden, 2011), the Australian National Data Service's Research Data Management in Practice manual (ANDS, 2013) and by the ICPSR archive's report mentioned above (2012) each do a more thorough job of stepping their audience through the research life cycle with best practices for managing data at each stage of the process.

Finally, outside of the data and research life cycles, a notable model that highlights the curation aspects of the data lifecycle is the DCC Curation LifeCycle Model (DCC, 2012), aimed at digital curators. The visual of the model is quite complex, but the stages and actions helped form the basis for our 2013 Data Curation Pilot project's draft curation workflow (described later in this report). Table 1 gives the full example of how the DCC model was adapted for the purposes of the pilot. The author was introduced to this model in 2011 in a Digital Curation 101 full-day workshop that introduces researchers and data custodians to the stages of the Curation LifeCycle Model.

Table 1: Stages of Data Curation, adapted from the DCC's Curation LifeCycle Model

Data Curation Stage	**Researcher Role**	**Curator Role**
Conceptualize	Write a Data Management Plan (DMP), plan the creation of the data, methodology, etc.	Train researchers on how to write a DMP.
Create	Capture data and document the process (include descriptive, structural, and technical metadata).	Develop tools to help researchers document their data and to capture metadata that will facilitate reuse.
Recruit	Be aware of collection policies of the repository. Approach repository with data that fall within those criteria.	Create clear documents and collection policies.

Receive	Deliver data with appropriate documentation, including metadata.	Develop interface for complete transfer of data from researcher that includes necessary metadata.
Appraise and Select		Evaluate data for long-term curation and preservation. Select appropriate data.
Ingest		Transfer data to appropriate archive, repository or other location.
Arrange and Describe		Collect and assign representation information. Determine the relationships between objects. Create metadata, using appropriate standards.
Preservation Actions		Move to a secure storage location. Transform file formats, authenticate, integrity check and/or include data cleaning, validation, ensuring acceptable data structures.
Dissemination and Access		Make data accessible by displaying publicly or by exposing metadata to other systems. Allow for access or download once discovered. Generate a permanent identifier for long-term citation.
Transform	Archive new versions.	Through reuse, versioning or migration.
Reappraise and/or dispose	Track reuse indicators.	Evaluate the impact or value of the data and determine whether to keep or dispose.

Note: Highlighted areas are the focus of the pilot project.

Archival Best Practices for Digital Curation

Best practices derived from archival disciplines can be extremely useful for data curators, but according to Tebo and Lee are often overlooked by developers of new data curation services in academic and disciplinary settings (2012). These workflows build on the best practices for archiving and curation physical collections, a deep body of knowledge that the archival profession has generated from decades of experience.

Faced with the changing nature of archival collections - imagine floppy disks instead of manila folders and hard drives stacked up next to banker boxes - the community has taken action to translate their skills from the analog to the digital. For example, the Society of American Archivists offers courses, both in-person and online, on the subject of digital archives (2013). The courses and the resulting Digital Archives Specialist (DAS) Certificate Program, which began in May 2011, are retooling the archival community with the knowledge and skills to apply their profession’s best practices to managing and curating digital objects. According to Esposito (2012) over 800 archivists have taken at least one DAS course in its first year. Similarly, the National Digital Stewardship Residency (NDSR, 2013) kicked of in 2013 with support from the Library of Congress and the Institute of Library and Museum

Services. NDSR provides a post-graduate experience to recent Master of Library and Information Science degree holders who seek to develop their digital stewardship skills in an immersive and hands-on way.[5]

Taking best practices from analog curation one step further, the archival community is researching and designing best practices for digital preservation and treatment of born-digital objects as well. This research is best illustrated by the OCLC report "Walk This Way" (Barrera-Gomez, 2013) and the University of Virginia-based AIMS Model for born-digital collections (AIMS Work Group, 2012). These models detail the step-by-step procedure to process a digital collection, including steps to quarantine the digital files to avoid virus exposure, use write-blockers to avoid contaminating metadata, and arrange and describe collections encompassing a variety of file types.

Practitioner Approaches to Data Curation

The UMN's project uses a "pilot" approach to understand and document the stages of data curation in order to draft a workflow for data curation in the Libraries. Our approach was loosely based on a similar project from the University of California - San Diego. Although the details of the UCSD project were not known at the time of drafting the 2013 Data Curation Pilot proposal, their findings and implementation documentation are now available (Minor, 2013). The UCSD pilot's approach was an in-depth look at five specific research groups, the way that they do research, and the data that they produce, in an effort to understand the full research life cycle of the data. Their resulting services may benefit from the detailed user-needs assessment; however, the scaling such an approach may be difficult to because of the time and effort that went into to crafting unique responses to each of their five pilot participants. In contrast, the UMN pilot sought to utilize existing tools and services in an effort to better understand our current capacities and limitations.

Practical curation approaches for digital data from institutional repository practitioners are mainly found in conference and web presentations. The University of Edinburgh's DataShare (Rice, 2013) and the Purdue University Research Repository (PURR; Mathews and Witt, 2013) are excellent examples of how academic institutions are handling the archival workflows for research data that were recently presented at the Open Repositories conference in Prince Edward Island. At this same conference, Humphry (2013) gave a presentation addressing the challenges of research data to traditional institutional repositories. Trident, a scientific workflow workbench developed by Microsoft Research, was presented at the 2012 Open Repositories as an industry response to the gap in tools for data curation (Kowalczyk & Plale, 2012).

Finally, within the context of disciplinary data repositories, several repository best practices rise to the top. One example is, NEESHub, a disciplinary repository for

[5] The author is a participant in the DAS program, with expected completion in 2014, and is a Curriculum Review panelist for the NDSR program.

earthquake engineering data, published workflows and holds web presentations on their curation techniques (Pejša and Hacker, 2013). DataOne holds research data from a variety of earth-observation related disciplines and publishes guides on data management. For example, their resources on ecological data (ESA, 2011) helps researchers prepare their data for eventual reuse in the archive. Finally, a very impressive example of detailed data curation comes from the oceanographic researchers. The Ocean Data Cookbook (Leadbetter, Raymond, Chandler, Pikula, Pissierssens, and Urban, 2013) describes step-by-step instructions for curating this type of data using a DSpace-based repository. The Cookbook was developed through use-cases, with digital object identifiers (DOIs) as a central component to the curation approach (Raymond, 2013).

References

AIMS Work Group. 2012. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf.

Australian National Data Service (ANDS). (2013). Research Data Management in Practice. http://www.ands.org.au/datamanagement/data-management-practice-guide.pdf

Barrera-Gomez, J. and Erway, R.. 2013. Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf.

Carlson, J. (2014). The Use of Life Cycle Models in Developing and Supporting Data Services. In J. M. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals* (pp. 36-39, review copy). Purdue University Press.

DataOne. (2013). Best Practices. http://www.dataone.org/best-practices

DDI Alliance. (2004). DDI Version 3 Conceptual Model. http://www.iassistdata.org/downloads/W5_DDI.pdf.

Digital Curation Center (DCC). (2013). DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model

ESA 2011 (2011). How to Manage Ecological Data for Effective Use and Re-use. DataOne. http://www.dataone.org/esa-2011-how-manage-ecological-data-effective-use-and-re-use

Esposito, J. R.. (2012). Digital Curation: Building an Environment for Success. In Duranti, L. and Shaffer, E. (Eds.), Proceedings of the International Conference "The Memory of the World in the Digital Age: Digitization and Preservation" (pp. 624-635). UNESCO. http://www.ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf

Humphrey, C.. (2006). e-Science and the Life Cycle of Research. http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

Humphrey, C.. (2013). The Long Tail of Data Wagging the Institutional Repository. Open Repositories 2013, Charlottetown, Prince Edward Island, Canada. http://or2013.net/sites/or2013.net/files/slides/OR2013_Workshop_Humphrey_0.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI. http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Kowalczyk, S.T. & Plale, B. (2012). Workflows for Digital Preservation and Curation Workshop. Presented at The 7th International Conference on Open Repositories (OR2012), Edinburgh, Scotland, July 2012. http://d2i.indiana.edu/sites/default/files/or12_trident_workshop_slides.pdf

Leadbetter, A., Raymond, L., Chandler, C., Pikula, L., Pissierssens, P., Urban, E.. (2013). Ocean Data Publication Cookbook. International Oceanographic Data and Information Exchange. IOC Manuals and Guides No. 64. http://www.iode.org/mg64

Matthews, C. E. and Witt, M.. (2013, July 8). The Purdue University Research Repository: Providing Institutional Data Services With a Virtual Research Environment, Data Publication and Archiving. Workshop: Institutional Repositories Dealing with Research Data, hosted by the DCC, IASSIST, and COAR. Open Repositories 2013, Charlottetown, Prince Edward Island, Canada. http://www.slideshare.net/courtneyearlmatthews/purr-or2013

Minor, D. (2013, April 4). UC San Diego Curation Pilots: Planning for the Future. 5th Annual University of Massachusetts and New England Area Librarian e-Science Symposium hosted by the University of Massachusetts in Shrewsbury, Mass. http://escholarship.umassmed.edu/escience_symposium/2013/program/6/

National Digital Stewardship Residency (NDSR). (2013). National Digital Stewardship Residency. Digital Preservation (Library of Congress). http://www.digitalpreservation.gov/ndsr/

Pejša, S. and Hacker, T.. (2013), Curation of Earthquake Engineering Research Data. NEESHub. http://nees.org/resources/5751.

Raymond, L. (2013, May 22). Publishing and Citing Ocean Data. OneNOAA Science Seminar, National Oceanographic Data Center. http://www.nodc.noaa.gov/seminars/2013/support/Lisa_Raymond_OneNOAASeminar_slides.pdf

Rice, R.. (2013, July 8). On Being a Cog Rather Than Inventing the Wheel: Edinburgh DataShare as a key service in the University of Edinburgh's RDM Initiative. Workshop: Institutional Repositories Dealing with Research Data, hosted by the DCC, IASSIST, and COAR. Open Repositories 2013, Charlottetown, Prince Edward Island, Canada. http://or2013.net/sites/or2013.net/files/slides/OR2013-Workshop-Rice_0.pdf

Rohrs, L., Conte, J., Guss, S., Smith, M., Rutkowski, A., and Mistry, H.. (2013). Data Management Planning and Services around the Data Life cycle at New York University. Unpublished white paper accessed on November 27, 2013 from https://rdmi.uchicago.edu/sites/rdmi.uchicago.edu/files/uploads/Rohrs,%20L.,%20Conte,%20J%20and%20Mistry,%20H_Data%20Management%20Planning%20and%20Services%20around%20the%20Data%20Life-Cycle%20at%20NYU_0.pdf

Society of American Archivists. (2013). Digital Archives Specialist (DAS) Curriculum and Certificate Program. http://www2.archivists.org/prof-education/das.

Tebo, H.R. and Lee, C. A.. Closing the Digital Curation Gap: A Grounded Framework for Providing Guidance and Education in Digital Curation. Archiving 2012 Final Program and Proceedings from the 2012 Society for Imaging Science and Technology. http://www.ils.unc.edu/callee/p57-tibbo.pdf

Tenopir, C., Birch, B., and Allard, S.. (2011). Academic Libraries and Research Data Services: Current Practices and Plans for the Future. ACRL White Paper. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

University of Illinois at Urbana-Champaign (UIUC) Graduate School of Library and Information Science (iSchool). (2013). Specialization in Data Curation. http://www.lis.illinois.edu/academics/programs/specializations/data_curation

University of Virginia Data Management Consulting Group. (2013). Steps in the Research Life Cycle. http://dmconsult.library.virginia.edu/lifecycle/

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., and Horton, L.. (2011). *Managing and Sharing Data, Third Edition*. UK Data Archive: University of Essex. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Background

The University Libraries offers data management services to our diverse campus community of nearly 70,000 faculty, staff, and students. Our current services include consultation on writing data management plans; training for faculty, staff, and students in creating quality digital research data; and tools for discovering and ethically reusing data.[6] These services were developed in 2009-2010 in response to an interest in data management on campus as well as a charge from the National Science Foundation that data management plans be included with all grant proposals submitted after January 18, 2011.[7]

In a separate initiative, more closely related to author's publishing rights and open access issues in scholarly communications, the Libraries created and launched an institutional repository, the University Digital Conservancy (UDC; https://conservancy.umn.edu) in 2007. The UDC provides free, worldwide access to scholarly and administrative works produced by or about the University of Minnesota. The UDC is open to self-deposit by UMN affiliates and currently has over 35,000 records that have been downloaded over 3,300,000 times (as of October

[6] A full list of services is available at http://www.lib.umn.edu/datamanagement.

[7] DMP guidelines by NSF available at http://www.nsf.gov/eng/general/dmp.jsp.

2013).[8] Datasets make up a small percentage of these assets by university researchers and students. However, there are several examples of datasets that were deposited to the University Digital Conservancy prior to the 2013 Data Curation Pilot. Some examples include:

- Aerospace Engineering Data Collection, http://purl.umn.edu/101457
- Astronomical Data Example, http://purl.umn.edu/116310 https://conservancy.umn.edu/handle/116310
- Historic Scientific Observational Data Example, http://purl.umn.edu/138532
- Institute for Health Informatics Data Collection, http://purl.umn.edu/132498
- Minnesota Geological Survey, GIS Supplements to born-digital maps, http://purl.umn.edu/708

In addition to the UDC, another data archiving tool provided by the University Libraries is the UMedia Archive for digital video, images, and audio files (http://umedia.lib.umn.edu). Here again, data may be found in the form of primary sources for the digital arts and humanities.

The University Libraries anticipates increased campus needs for both data management and repository services, given user-needs assessments and directions of federal funders. Data curation as a library service was included as a component of the Research Services Framework, drafted by a subgroup of the University Libraries Research & Learning Directors on January 3, 2013. Shortly before the Data Curation Pilot project was initiated, several unpublished library reports were reviewed and incorporated into the data curation component of the framework; they include:

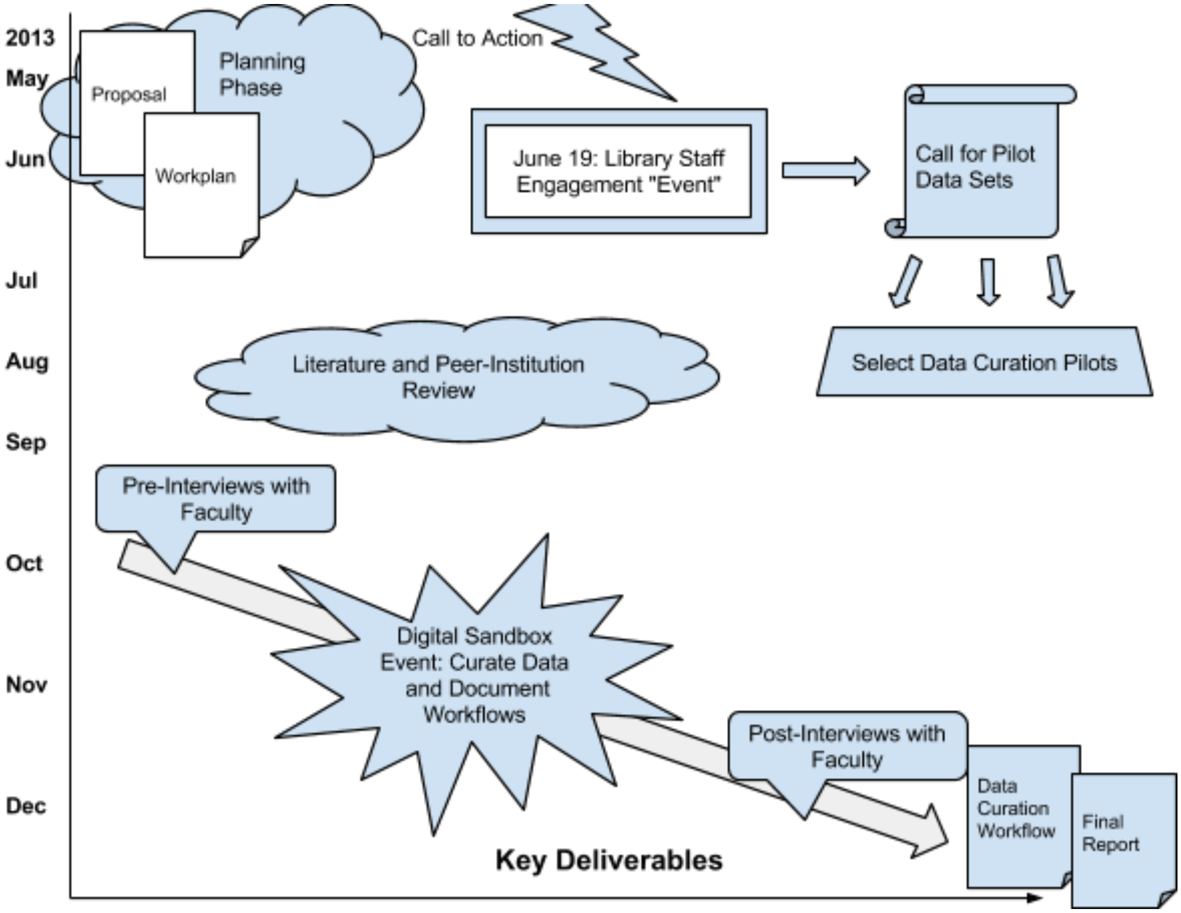
- 2012: *An Agenda for Deepening Library Support for Research*, by John Butler, Layne Johnson, and Lisa Johnston. This capstone project was written in participation with the ARL E-science Institute, 2011-12 Cohort. One of its primary recommendations is to implement data curation services for the campus.
- 2011: *Near-Term Recommendations for Action from the Data Management, Access, and Archiving Working Group* (DaWG; Subgroup of the Libraries Research Support Services Collaborative) by Stephen Hearn, Kristi Jensen (co-chair), Lisa Johnston (co-chair), Meghan Lafferty, Jon Nichols, Beth Petsan, and Amy West. DaWG explored service models and campus needs related to data and identified several opportunities for investment from the Libraries.
- 2009: *Data Stewardship Opportunities for the University of Minnesota Libraries: Recommendations from the E-Science Data Services Collaborative (EDSC)* by Tony Fang, Gary Fouty, Cody Hanson, Amy Hribar, Kristi Jensen (co-chair), Lisa Johnston, Peter Kirlew (co-chair), Wayne Loftus, Jon Nichols, and Amy West. Outcomes of this report includes the creation of the Data Management web page (http://www.lib.edu/datamanagement) and the creation of the library's educational approaches to data management.

[8] See current download stats of the UDC at https://conservancy.umn.edu/stats_display.jsp?handle=1

Methodology

The Data Curation Pilot project was completed over an eight-month period (May-December) in 2013. The project was implemented by the author through several phases, as illustrated in the visual roadmap in Figure 2. The actions taken in each of the five project phases are detailed in this section.

Figure 2: Visual Roadmap for the 2013 Data Curation Pilot Project


2013
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Proposal
Workplan
Planning Phase
Call to Action
June 19: Library Staff Engagement "Event"
Call for Pilot Data Sets
Select Data Curation Pilots
Literature and Peer-Institution Review
Pre-Interviews with Faculty
Digital Sandbox Event: Curate Data and Document Workflows
Post-Interviews with Faculty
Data Curation Workflow
Final Report
Key Deliverables

Phase 1: Engagement (May-December 2013)

A goal of the pilot project was to engage with campus partners in issues surrounding data management. Data service providers on campus are growing. For example, in the summer of 2013, the College of Liberal Arts hired a Data Management Specialist to engage with researchers in the college. As part of the PEL program (which has an emphasis on change management and engagement), the Data Curation Pilot set out to engage stakeholders in informal and formal conversations regarding data management issues, as well as implement change management techniques in the Libraries regarding our new role in data curation. The pilot completed the following activities:

- Met with key stakeholders in the library early in the project – including University Libraries Cabinet, library directors, project sponsors, and several library units – and shared a work plan to move forward.
- Held a Libraries staff engagement event June 19, 2013 that brought in approximately 40 staff members to hear a presentation on the Data Curation Pilot and see examples of how the Libraries is currently handling digital data collections.
- Held a campus-wide research data management discussion event on June 27, 2013, that brought together 21 data service providers (from Office of information Technology (OIT), Minnesota Supercomputing Institute (MSI), Academic Health Center (AHC), College of Liberal Arts (CLA), etc.) and kick-started an informal community of practice (iCoP) for Research Data Management on campus. We continue to meet monthly and present on topics such as metadata standards, spatial data management issues, best practices, etc.
- Presented the pilot project to Brian Herman, the Vice President for Research, and colleague PEL Circle members on August 22, 2013, due to Dr. Herman's role as a 2013 PEL mentor.
- Presented pilot results to campus at the PEL Closing celebration event held on December 3, 2013.

Phase 2: Identifying Data for the Pilot (July-September 2013)

This phase of the project included establishing the selection criteria of the pilot data, promoting the call for proposals, analyzing the response from campus researchers, and selecting the datasets to be included in the pilot.

Selection Criteria

At the onset, this project recognized that in order to be successful, the datasets selected for the pilot must match certain criteria as defined by our current capacities. Therefore, the call for proposals[9] identified several selection criteria for the data. These criteria were:

- Public access: Curated data in the pilot will be released for public access via the web (e.g., Digital Conservancy, UMedia).
- Availability: The data should be complete and ready for public distribution by September 2013.
- Restrictions: Restricted access datasets will not be selected in the pilot. The data will be openly accessible to the public and therefore must not contain any private, confidential, or other protected information. (Note: This item linked to University Policies on Privacy and Data Security[10] and examples of private data[11].)

[9] The web version of the call for proposals is at https://z.umn.edu/datapilot13, and the form used to capture responses to the call is in Appendix B.

[10] See http://www.privacysecurity.umn.edu/policies/home.html

[11] See http://www.policy.umn.edu/Policies/Operations/OPMisc/INTERNALACCESS_APPB.html

- Authorship: The data must be authored, produced, and/or sponsored by a UMN faculty member or researcher who is willing and able to grant the deposit agreement[12] for UMN to preserve and distribute the data.
- Size: The data may contain multiple files. However, individual files cannot exceed 1GB per file or the UDC's current upload setting.
- Documentation for reuse: The data must include adequate documentation describing the nature of the data at an appropriate level for reuse and discovery.
- Variety: Datasets should reflect a variety of data types, formats, and disciplines to account for the range of possible data curation workflows and considerations.
- Time commitment: Faculty participants must be willing to partner with the library in a "pilot" atmosphere and contribute at least two hours of their time over the course of the Fall 2013 semester to participate in pre- and post-curation interviews.

Response to the Call for Proposals

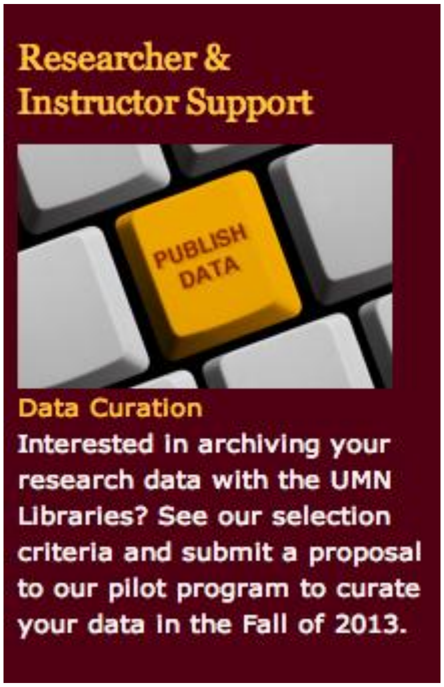
The pilot's call for proposals was issued in July 2013 via several communications vehicles, including the Library's homepage (Figure 3), an email from library subject liaisons to the majority of UMN department faculty and staff, Twitter (@walterlibrary), Facebook (Walter Library), the author's work email signature byline, and the July 15, 2013 edition of the bi-weekly UMN Graduate Brief. Distribution was a success, as evident from the statistics of the z.umn link used in the call, http://z.umn.edu/datapilot13, which was visited 457 times from July 2nd - November 27, 2013.

The call resulted in 16 proposals submitted to the Data Curation Pilot. A summary of the responses received by collegiate unit is in Figure 4, and the breakdown of how the responses meet the selection criteria is presented in Table 2. Additionally, respondents were asked to "Tell us why you are interested in helping the UMN Libraries explore and pilot 'data curation services' which includes archiving, preservation, and access – ultimately with reuse in mind." The responses to the qualitative component of the survey reinforced the observed user-need for data curation services. For example:

- "I recognize that I'm not the only person in this predicament of storing larger sets of data (conceived broadly) and that figuring out how to do this well and sustainably will help many, many folks around the University."
- "With data management plans required by NSF, participating in such a project would help us satisfy this requirement. Additionally, means to store and curate our data would allow us to do more with it over time (i.e., go back to it and reanalyze/reassess) which could lead to new discoveries."
- "Data curation goes beyond backup and storage. Meanwhile, how to archive, preserve, and provide access to (sometimes large) datasets is still new to many researchers."

[12] See the UDC deposit agreement at https://conservancy.umn.edu/basicdeposit.pdf

Figure 3: Promotional Example for Submitting a Proposal to the 2013 Data Curation Pilot (UMN Libraries Homepage)

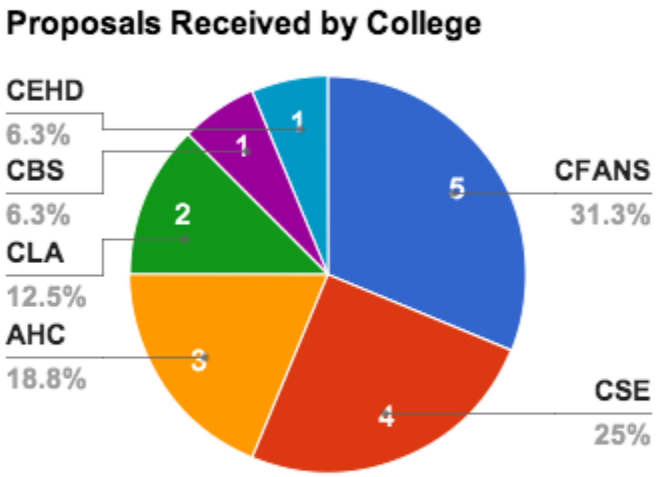


Researcher & Instructor Support

Data Curation

Interested in archiving your research data with the UMN Libraries? See our selection criteria and submit a proposal to our pilot program to curate your data in the Fall of 2013.

Figure 4: Results of the Call for Proposals to the 2013 Data Curation Pilot as Displayed by College (n=16).



Legend

- **College of Food, Agricultural and Natural Resource Sciences (CFANS)**
- **College of Science and Engineering (CSE)**
- **Academic Health Center (AHC)**
- **College of Liberal Arts (CLA)**
- **College of Biological Sciences (CBS)**
- **College of Education and Human Development (CEHD)**

Table 2: Results of the Call for Proposals to the 2013 Data Curation Pilot (n=16).

Criteria	Results (n=16)	Notes
Variety	Excel (8), Raw (4), image (3), SPSS (2), video (2), GIS (1), papers (1)	Datasets should reflect a variety of data types, problems, and disciplines.
Public Access	88% (14/16) agree, "Archiving, preservation, and public access are important."	But, 38% (6/16) "do not necessarily need public access to [their] data"
Availability	100% (16/16)	Data must be ready by Sep 2013.
Restrictions	19% (3/16) contain data restrictions.	Such as private, confidential, or protected information.
Authorship	19% (3/16) unsure about the ownership rights to their data	Necessary to grant the deposit agreement.
Size	25% (4/16) have data files greater than 1GB.	Answers include "occasionally greater" or "not sure."
Documentation for Reuse	38% (6/16) do not have adequate documentation to facilitate reuse but are interested in assistance.	Documentation describes the nature of the data and provides an appropriate level of context for reuse.

Datasets Selected for the Pilot

Initially, we selected five datasets for the pilot. However, two additional datasets were invited after two of the initial authors were unable to deliver their data within the Fall 2013 timeframe. Ultimately, five datasets were successfully received by the faculty authors and included in the pilot. These were:

- Engineering Data: GIS data from reverse-engineering print transportation maps created by David Levinson (Civil Engineering, CSE).
- Health Sciences Data: Excel and .csv data from periodontal clinical trials created by James Hodges (School of Public Health, ACH).
- Interdisciplinary Data: Excel data of chemical traces found in Minnesota lakes created by Bill Arnold (Civil Engineering, CSE).
- Natural Resources Data: SPSS data from online tourism surveys created by Lisa Qian (Forest Resources, CFANS and Extension).
- Social Sciences/Humanities Data: Video and transcription files from Ojibwe conversations created by Mary Hermes (Curriculum and Instruction, CEHD).

The actual data received from each author, as well as a summary of the pre-curation interview (described below), is included in Appendix C.

Phase 3: Interviews with Researchers (September-October 2013)

Between September 19th and October 7, 2013, the author scheduled and met with the pilot researchers, accompanied by the subject librarian to that discipline when available, to conduct the pre-curation interviews (in person, via phone, or through Skype). The interviews set expectations for the pilot, reviewed any potential curation concerns (privacy, file size), and discussed the researchers' individual archiving needs. The primary aim of these discussions was to narrow down the scope of the data suggested for inclusion in the pilot (in some cases, multiple projects' worth of data) and to identify any relevant documentation about the data creation that would be necessary for reuse. Follow-up instructions were emailed to the faculty after each interview to outline the actions needed on their part:

1. Deliver the data and any accompanying documentation to the Libraries via email, Google Docs, or Dropbox;
2. Fill out a metadata form for the dataset, supplied by the author as a list of elements adapted from the Dublin Core schema; and
3. Sign and return a UDC deposit agreement.

The five datasets and the accompanying materials were delivered to the library by November 1, 2013. It should be noted that even with the clear criteria for involvement, the first interview several dataset qualities that were not in line with our project's requirements. For example, some researchers were unaware of or underestimated the actual size of their data. When the conversations dove a bit deeper into proprietary issues and ownership concerns, two authors discovered that their data might not be publicly shareable. One researcher had very restrictive IRB agreements for a dataset, even though the dataset did not contain human subject data, but rather included personally identifiable information that could be removed. The researcher explained that a colleague in the psychological sciences wrote the IRB agreement, as this was not something commonly done in the primary authors' field.

Overall, faculty were less concerned with archiving their data for others to access, or even to meet federal mandates. They each wanted a permanent home for the data to "live on," possibly only for them to access. The pilot was one way of ensuring this, with the bonus of open accessibility.

Phase 4: Develop Curation Workflows and Archive the Datasets (November 2013)

To determine the appropriate steps and workflow to curate our data in the pilot, the project included an engagement event to brainstorm and share knowledge on this topic of digital data curation. This event, called the Digital Curation Sandbox, took place on November 4, 2013, and brought together 23 staff members from across the libraries (as well as the CLA data management specialist) to share and build from our collective expertise. Staff were grouped into five teams, and each were assigned a dataset from the pilot to evaluate. The teams were comprised of a librarian liaison to the subject discipline of the dataset, an archivist or curator from the libraries'

Archives and Special Collections unit, and a staff member from the Libraries Digital Technology unit who brought either a cataloging and metadata perspective or an IT perspective. Finally, a "data facilitator" led each team through the event's exercises. The "data facilitator" was an expert in the data type and format of that group (for example, the director of the Borchert Map Library led the group looking at GIS data files).

As pre-work for the digital sandbox, the participants read a summary handout of the dataset that they would be working with. These handouts are available in Appendix C, and each contains a brief overview of the faculty member, their dataset, the metadata contributed by the author, and a Submission Information Packet (SIP) that presented each of the files delivered to the library along with select screenshots of the data. (Note: The data itself was distributed at the event.) Since the Digital Curation Sandbox was held in an active learning classroom, each team had at their disposal a whiteboard, computer hookups for each participant, and a dedicated monitor. The agenda for the event (Appendix D) included three activities:

1. Engage with a conceptual digital curation workflow to develop a shared terminology,
2. Apply participant skills and expertise to the datasets in the pilot, and
3. Wrap up with a discussion around a common workflow for curating digital data in the library.

As a result of the first activity, the group determined the following stages of digital curation that would be used as a shared terminology for the rest of the sandbox, shown in Table 3. In a notable difference from the Digital Curation Center's workflow stages, our group broke *Arrange and Describe* into two steps: *Organize* and *Description and Metadata*. This is most likely due to the familiar tools that we use in the library, such as our institutional repository (running on DSpace software), which may not allow for arrangement post-ingest.

Table 3: Stages of Data Curation as Described by UMN Staff Participants at the Digital Curation Sandbox

Group	Stage 0:	Stage 1:	Stage 2:	Stage 3:	Stage 4:	Stage 5:	Stage N:
UMN Digital Curation Sandbox	**Receive**	**Appraise / Inventory**	**Organize**	**Treatment Actions / Processing**	**Description / Metadata**	**Access**	**Reuse**

As a result of this planning, the datasets were curated according to best practices identified by library staff, as well as through the literature review of existing tools and technologies. The author met with the five sandbox group facilitators in separate one-hour meetings on November 22, 2013. Each facilitator brought their unique expertise to the data (type, software to use, etc.), and the session was used to walk once more through the draft data curation workflow and take the necessary actions to curate the data for reuse. That day, four of the five datasets were archived into the UDC collection, "Data Curation Pilot Project, 2013" (http://purl.umn.edu/160292). The final dataset needed additional processing and was released for access on December 2, 2013. Detailed overviews of the experiences and outcomes for each of the datasets are included in the Results section of this report.

Phase 5: Reflection and Assessment of Project (December 2013)

The final results of the pilot were delivered to the researchers for their feedback on December 2, 2013. Along with a link to their curated dataset, the author delivered a summary report to each data owner outlining the sets taken to curate their data, along with the draft curation workflow generated from the pilot. Next, a self-reflection survey was delivered to the five faculty authors on December 2, 2013. The survey included four questions about the data curation pilot. These questions were:

- What are 3 things that you find successful with the result of your curated dataset?
- What are 3 things that you consider lacking with the result of your curated dataset?
- Any additional comments for us?

The responses to our follow-up survey are included in Table 4 and reflect a generally positive assessment of how the data curation pilot handled their data.

Table # 4: Results of the Post-Curation Survey Delivered to Five Faculty Submitters (n=3).

Faculty Participant	Successes	Shortcomings	Additional Comments
Faculty #1	1.Knowing that there is a permanent link to the data and that those interested in this subject/topic can have access to the data and its documentation. 2.A non-proprietary version of the data! 3.As the survey is an ongoing effort, the Tourism Center will keep data curation in mind during future survey implementations	In the data dictionary, the values of each variable are listed separately from it name & label (etc.). I know that this is really hard to do, so I do not mean to complain about it :-)	I like the way the abstract was constructed. I figured that the abstract is actually my answers to several questions asked during the process. Answering questions feels less challenging than writing an abstract from scratch. I hope that one can click on the citation link (Tourism Center website in this case) and go directly to the website on which reports using the data can be found.
Faculty #2	"It got set up We didn't have to do much work Everything seems to be there and working. It would be nice if somebody downloaded it, but that's not up to you	(None.)	Nope. I'm not just going thru the motions here. The whole thing went very smoothly from my point of view. I don't know what to say. All I've ever wanted re this dataset was to make the dataset available, per

	all."		our U01 agreement with NIH and my general belief that scientific data (especially taxpayer-funded) should be as broadly available as possible. You've helped us do that.
Faculty #3	That data are permanently archived. They are easy to access. The website is clear as to what the data are.	Some expertise is still required to navigate them. The thesis is not yet linked (but could be once embargo is lifted). Not clear how people will find the data (i.e., will Google searching find it?) or how to announce its availability.	I think this will be a very important tool. With "Data Management" being critical in NSF proposals (and likely other funding agencies), it would be great to have a centralized service like this to which data could be submitted for curation, storage, and public access. Such access may be required of federally funded projects, so this is a big need. Something centralized (vs. each PI putting data in netfiles or on departmental web servers) is definitely needed.
Faculty #4	Got the big hand over [of the data] started. Established a process for continuing to add data in the future. Started a relationship with the Librarians (archives).	Just my fault, have [many more data files] still to hand over.	I could really use structure to continue to hand things over, I wonder if we could set up a timeline - as the formal grant project is over. This is the crucial (but not immediate) part!
Faculty #5	The data is permanently archived. The data is available online. The data is findable via search engine.	Only includes 1958 map. Other data was provided The data is only as well-documented as we provided. It would be nice if documentation could be (magically?) enhanced. Abstract/descriptions should have hyperlinks to reports, which use (more fully describe) datasets.	Online mapping/GIS, i.e. tie datasets to real world mapping software. Maybe work with NHGIS project at the Minnesota Population Center.

Note: Responses are ordered in the way they were received.

Researchers were also asked what they would like to see in future services offered by the library. These responses reflected the broader campus need for data curation and individuals' disciplinary and data-specific needs. The responses were:

- I hope that one can click on the citation link (Tourism Center website in this case) and go directly to the website on which reports using the data can be found.
- I don't know what to say. All I've ever wanted re this dataset was to make the dataset available, per our agreement with NIH and my general belief that scientific data (especially taxpayer-funded) should be as broadly available as possible. You've helped us do that.
- With "Data Management" being critical in NSF proposals (and likely other funding agencies), it would be great to have a centralized service like this to which data could be submitted for curation, storage, and public access. Such access may be required of federally funded projects, so this is a big need. Something centralized (vs. each PI putting data in netfiles or on departmental web servers) is definitely needed.
- Online mapping/GIS, i.e. tie datasets to real world mapping software. Maybe work with NHGIS project[13] at the Minnesota Population Center.

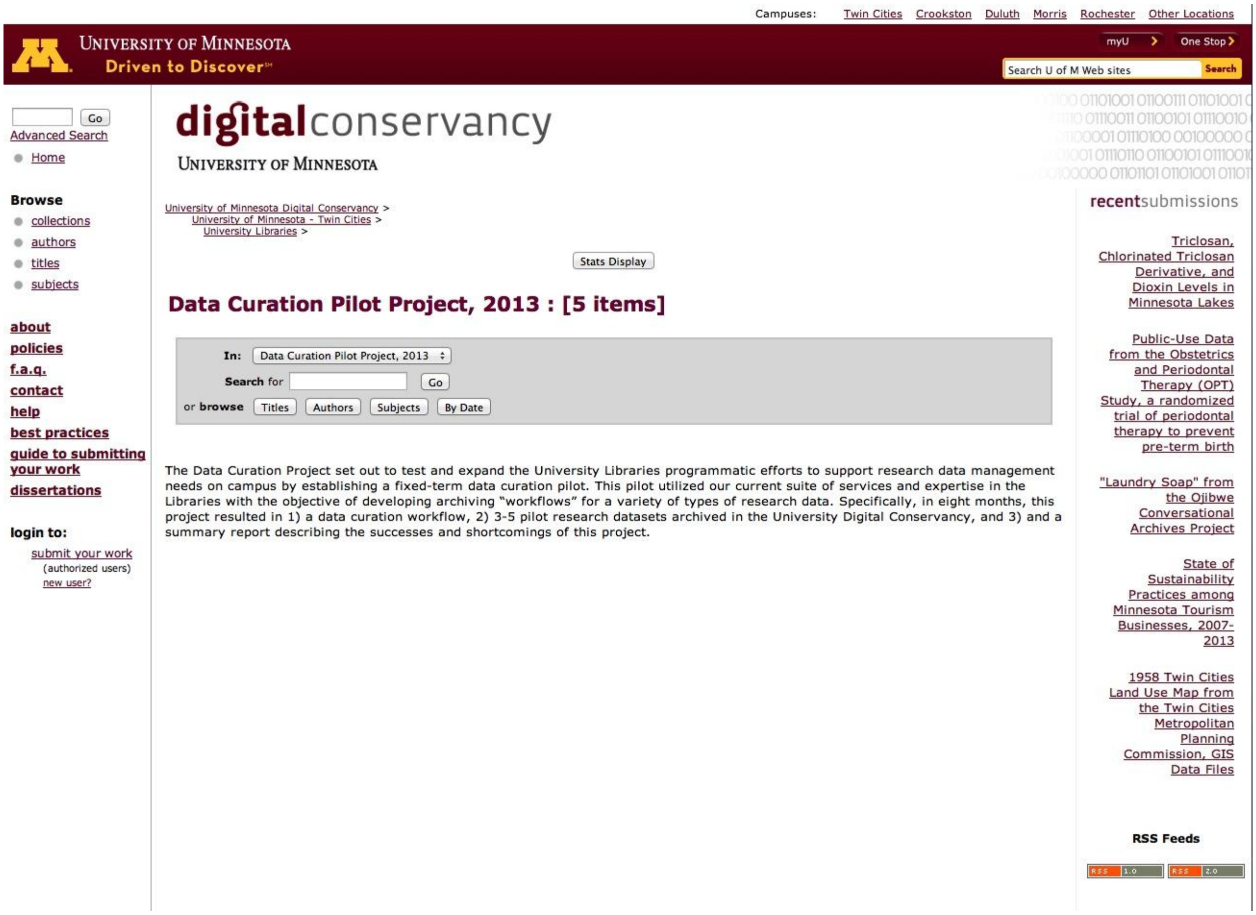
Results of the Pilot

As a result of the data curation workflow outline in the libraries' Digital Curation Sandbox event, and through the in-depth follow-up conversations with staff, the libraries public curated the five datasets for access and reuse into the University Digital Conservancy (UDC), UMN's institutional repository for open access to university research and archival materials. The data are included in a collection called the "Data Curation Pilot Project, 2013" at http://purl.umn.edu/160292 (Figure 5).

In this section, each dataset's treatment will be detailed. Next, the generalizable workflow for the entire curation pilot is outlined, based on the experiences for each data type and through the combined expertise of the staff involved in the project.

[13] The National Historical Geographic Information System (NHGIS) provides, free of charge, aggregate census data and GIS-compatible boundary files for the United States between 1790 and 2012. See https://www.nhgis.org.

Figure 5: Screenshot of the Data Curation Pilot Project, 2013 collection in the University Digital Conservancy



Data Curation Pilot Project, 2013 : [5 items]

The Data Curation Project set out to test and expand the University Libraries programmatic efforts to support research data management needs on campus by establishing a fixed-term data curation pilot. This pilot utilized our current suite of services and expertise in the Libraries with the objective of developing archiving "workflows" for a variety of types of research data. Specifically, in eight months, this project resulted in 1) a data curation workflow, 2) 3-5 pilot research datasets archived in the University Digital Conservancy, and 3) and a summary report describing the successes and shortcomings of this project.

recentsubmissions

Triclosan, Chlorinated Triclosan Derivative, and Dioxin Levels in Minnesota Lakes

Public-Use Data from the Obstetrics and Periodontal Therapy (OPT) Study, a randomized trial of periodontal therapy to prevent pre-term birth

"Laundry Soap" from the Ojibwe Conversational Archives Project

State of Sustainability Practices among Minnesota Tourism Businesses, 2007-2013

1958 Twin Cities Land Use Map from the Twin Cities Metropolitan Planning Commission, GIS Data Files

Curation Treatments for Each of the Five Pilot Datasets

The treatment for each type of data was captured in the Digital Curation Sandbox using a Google Docs spreadsheet (See example in Appendix E). The treatment actions for each dataset are described below.

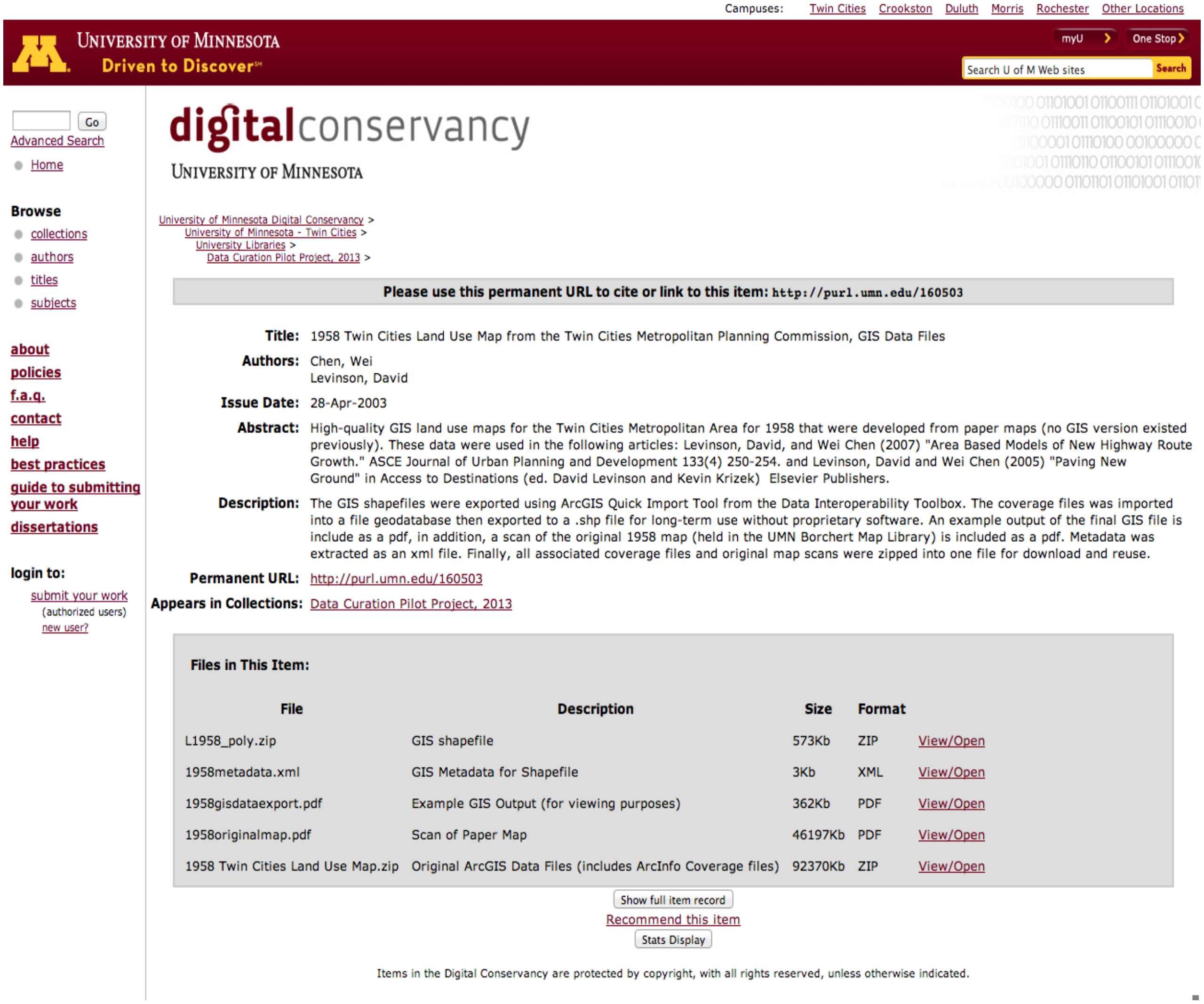
Engineering Data

This dataset (Figure 6) included GIS data from reverse-engineering print transportation maps created by David Levinson (Civil Engineering, UMN) and is available at http://purl.umn.edu/160503. The original submission information (see appendix C.1) was curated in the following way:

Data Curation Stage	Actions Taken With This Dataset
0. Receive	● Data files received on 10-09-13. The deposit agreement was received (PDF) on 10-16-13. ● Appropriate archive: It was determined that the appropriate home for this data would be the UDC. However, USpatial,

	though not yet ready as an archive for GIS data, will be a good choice in the future. ● Possible ownership concern: the original 1958 scanned map, authored by a state agency. This was determined “Low risk” as the maps were scanned from the UMN Library collection and the stage agency in question is very interested in getting their publications more openly available.
1. Appraise and Inventory	● Determine if the spatial data format(s) contain only proprietary data (ESRI ArcGIS) or include the more interoperable shapefiles (.shp). ● Identify the important files: in this case, the folder "L1958" in the "GIS" folder. These are the .shp files. ● Identify the metadata files (XML) in FGDC format. ● Missing information: The attribute table for the Landuse codes needs to be updated to define codes. Contact author.
2. Organize	● Understand the file structure, very complex in this case. ● Determine which files should be archived with the final datasets. In this case several of the scanned maps were duplicated as versions that were hard to distinguish by the file name. There were left as is.
3. Treatment Actions	● Create a final GIS output of the map as an image file. ● Preservation: ○ Convert the scanned map (psd file) to PDF file for ease of access. ○ Export the L1958 coverage files into a file geodatabase for the interoperable shapefiles. ● Zip the original ESRI GIS files
4. Description and Metadata	● Consider granularity of GIS formats -- do all included items need to be described? ● Document the related files as a separate text file. ● Expose metadata for shapefiles as XML (outside of the zip) for full-text indexing. ● Create a brief description of the processing done on the files. ● Map author-submitted metadata to our metadata schema.
5. Access	● Upload the 5 files to the UDC: these are ○ GIS shapefile (zip) ○ GIS Metadata for Shapefile (XML) ○ Example GIS Output (for viewing purposes) (PDF) ○ Scan of Paper Map (PDF) ○ Original ArcGIS Data Files (includes ArcInfo Coverage files) (zip) ● Plan for future changes in GIS data formats.
N. Reuse	● Notify the author of the dataset availability, the persistent URL location, and the recommended citation for reuse by others. ● Author is encouraged to track download statistics.

Figure 6: Screenshot of the resulting UDC record for the Engineering Data Example available at http://purl.umn.edu/160503.



University of Minnesota Digital Conservancy >
University of Minnesota - Twin Cities >
University Libraries >
Data Curation Pilot Project, 2013 >

Please use this permanent URL to cite or link to this item: `http://purl.umn.edu/160503`

Title: 1958 Twin Cities Land Use Map from the Twin Cities Metropolitan Planning Commission, GIS Data Files

Authors: Chen, Wei
Levinson, David

Issue Date: 28-Apr-2003

Abstract: High-quality GIS land use maps for the Twin Cities Metropolitan Area for 1958 that were developed from paper maps (no GIS version existed previously). These data were used in the following articles: Levinson, David, and Wei Chen (2007) "Area Based Models of New Highway Route Growth." ASCE Journal of Urban Planning and Development 133(4) 250-254. and Levinson, David and Wei Chen (2005) "Paving New Ground" in Access to Destinations (ed. David Levinson and Kevin Krizek) Elsevier Publishers.

Description: The GIS shapefiles were exported using ArcGIS Quick Import Tool from the Data Interoperability Toolbox. The coverage files was imported into a file geodatabase then exported to a .shp file for long-term use without proprietary software. An example output of the final GIS file is include as a pdf, in addition, a scan of the original 1958 map (held in the UMN Borchert Map Library) is included as a pdf. Metadata was extracted as an xml file. Finally, all associated coverage files and original map scans were zipped into one file for download and reuse.

Permanent URL: http://purl.umn.edu/160503

Appears in Collections: Data Curation Pilot Project, 2013

Files in This Item:

File	Description	Size	Format	
L1958_poly.zip	GIS shapefile	573Kb	ZIP	View/Open
1958metadata.xml	GIS Metadata for Shapefile	3Kb	XML	View/Open
1958gisdataexport.pdf	Example GIS Output (for viewing purposes)	362Kb	PDF	View/Open
1958originalmap.pdf	Scan of Paper Map	46197Kb	PDF	View/Open
1958 Twin Cities Land Use Map.zip	Original ArcGIS Data Files (includes ArcInfo Coverage files)	92370Kb	ZIP	View/Open

Items in the Digital Conservancy are protected by copyright, with all rights reserved, unless otherwise indicated.

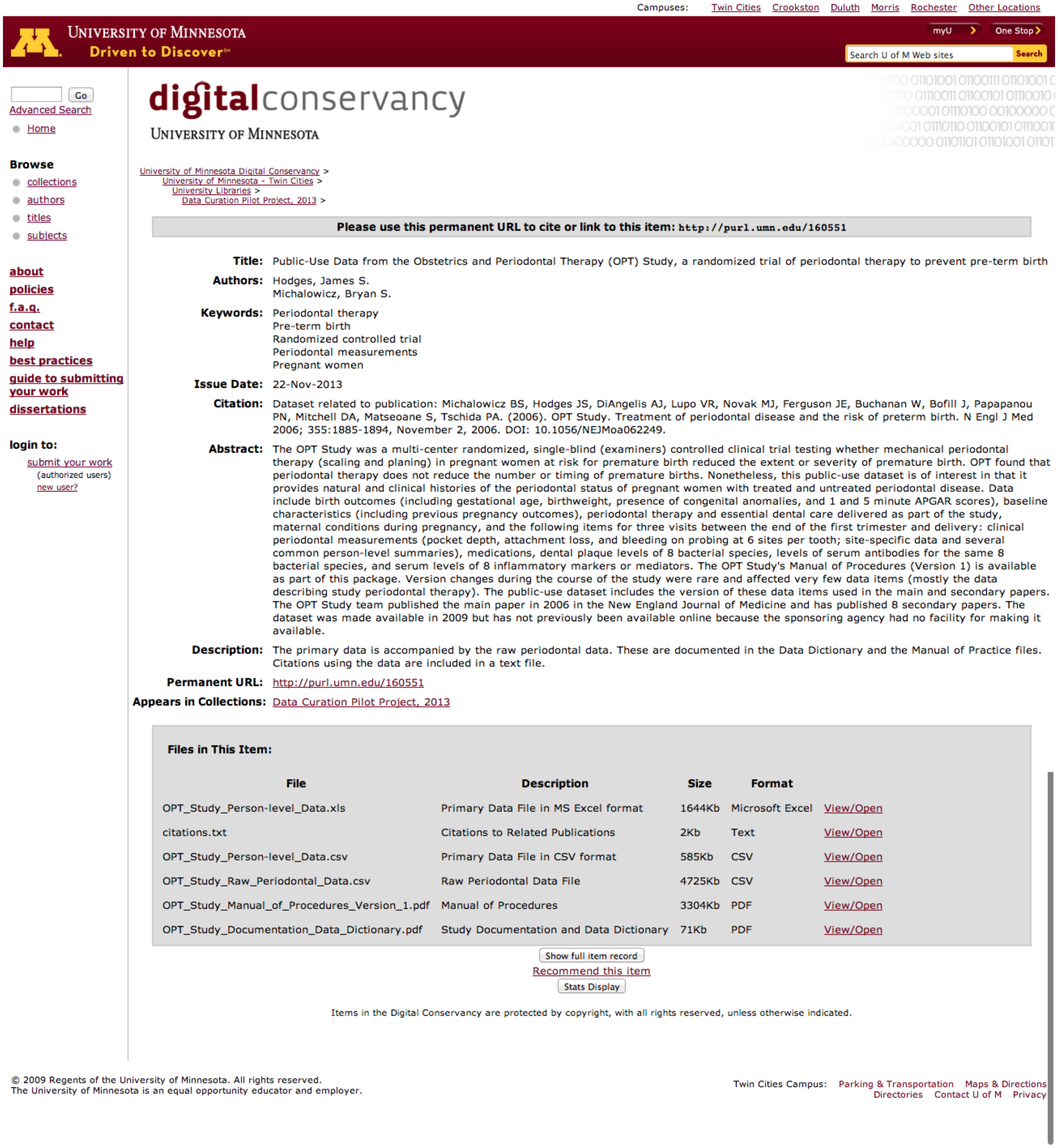
Health Sciences Data

This dataset (Figure 7) included Microsoft Excel and .csv data from periodontal clinical trials created by James Hodges (School of Public Health, UMN) and is available at http://purl.umn.edu/160551. The original submission information (see appendix C.2) was curated in the following way:

Data Curation Stage	Actions Taken With This Dataset
0. Receive	● Data files received on 09-26-13. The deposit agreement was received (online) on 09-26-13. ● Private/sensitive information: the files were checked to verify

	that they have been de-identified prior to receiving. ● Appropriate archive: It was determined that the appropriate home for this data would be the UDC.
1. Appraise and Inventory	● Documentation file was noted to be excellent. ● Inventory: Why are there 2 different raw data files? Was .txt file generated from .csv file or vice versa? Are they exactly the same?
2. Organize	● Files are pretty straightforward and the organization that we received is good. ● Did not include the .txt version of the data as this would be easily recreated from the .csv.
3. Treatment Actions	● Preservation: ○ Transform the excel file to .csv but keep both files as there will be loss of formatting in the csv. ○ Convert the two Word Docs to a PDF. This file included section breaks and therefore the Save as PDF function resulted in three PDFs that were then combined using Adobe Acrobat Pro.
4. Description and Metadata	● Map author-submitted metadata to our metadata schema. ● Citations to the main article was included in the Dublin core metadata. The citations to related publications were included in the author-submitted metadata and included as a .txt file.
5. Access	● Upload the five files to the UDC: ○ Primary Data File in MS Excel format ○ Primary Data File in CSV format ○ Raw Periodontal Data File (CSV) ○ Manual of Procedures (PDF) ○ Study Documentation and Data Dictionary (PDF) ○ Citations to Related Publications (TXT) ● The PDFs and text files will be keyword searchable providing more comprehensive access to the data.
N. Reuse	● Notify the author of the dataset availability, the persistent URL location, and the recommended citation for reuse by others. ● Author is encouraged to track download statistics.

Figure 7: Screenshot of the resulting UDC record for the Health Sciences Data Example available at http://purl.umn.edu/160551.



Please use this permanent URL to cite or link to this item: `http://purl.umn.edu/160551`

Title: Public-Use Data from the Obstetrics and Periodontal Therapy (OPT) Study, a randomized trial of periodontal therapy to prevent pre-term birth

Authors: Hodges, James S.
Michalowicz, Bryan S.

Keywords: Periodontal therapy
Pre-term birth
Randomized controlled trial
Periodontal measurements
Pregnant women

Issue Date: 22-Nov-2013

Citation: Dataset related to publication: Michalowicz BS, Hodges JS, DiAngelis AJ, Lupo VR, Novak MJ, Ferguson JE, Buchanan W, Bofill J, Papapanou PN, Mitchell DA, Matseoane S, Tschida PA. (2006). OPT Study. Treatment of periodontal disease and the risk of preterm birth. N Engl J Med 2006; 355:1885-1894, November 2, 2006. DOI: 10.1056/NEJMoa062249.

Abstract: The OPT Study was a multi-center randomized, single-blind (examiners) controlled clinical trial testing whether mechanical periodontal therapy (scaling and planing) in pregnant women at risk for premature birth reduced the extent or severity of premature birth. OPT found that periodontal therapy does not reduce the number or timing of premature births. Nonetheless, this public-use dataset is of interest in that it provides natural and clinical histories of the periodontal status of pregnant women with treated and untreated periodontal disease. Data include birth outcomes (including gestational age, birthweight, presence of congenital anomalies, and 1 and 5 minute APGAR scores), baseline characteristics (including previous pregnancy outcomes), periodontal therapy and essential dental care delivered as part of the study, maternal conditions during pregnancy, and the following items for three visits between the end of the first trimester and delivery: clinical periodontal measurements (pocket depth, attachment loss, and bleeding on probing at 6 sites per tooth; site-specific data and several common person-level summaries), medications, dental plaque levels of 8 bacterial species, levels of serum antibodies for the same 8 bacterial species, and serum levels of 8 inflammatory markers or mediators. The OPT Study's Manual of Procedures (Version 1) is available as part of this package. Version changes during the course of the study were rare and affected very few data items (mostly the data describing study periodontal therapy). The public-use dataset includes the version of these data items used in the main and secondary papers. The OPT Study team published the main paper in 2006 in the New England Journal of Medicine and has published 8 secondary papers. The dataset was made available in 2009 but has not previously been available online because the sponsoring agency had no facility for making it available.

Description: The primary data is accompanied by the raw periodontal data. These are documented in the Data Dictionary and the Manual of Practice files. Citations using the data are included in a text file.

Permanent URL: http://purl.umn.edu/160551

Appears in Collections: Data Curation Pilot Project, 2013

Files in This Item:

File	Description	Size	Format	
OPT_Study_Person-level_Data.xls	Primary Data File in MS Excel format	1644Kb	Microsoft Excel	View/Open
citations.txt	Citations to Related Publications	2Kb	Text	View/Open
OPT_Study_Person-level_Data.csv	Primary Data File in CSV format	585Kb	CSV	View/Open
OPT_Study_Raw_Periodontal_Data.csv	Raw Periodontal Data File	4725Kb	CSV	View/Open
OPT_Study_Manual_of_Procedures_Version_1.pdf	Manual of Procedures	3304Kb	PDF	View/Open
OPT_Study_Documentation_Data_Dictionary.pdf	Study Documentation and Data Dictionary	71Kb	PDF	View/Open

Twin Cities Campus: Parking & Transportation Maps & Directions
Directories Contact U of M Privacy

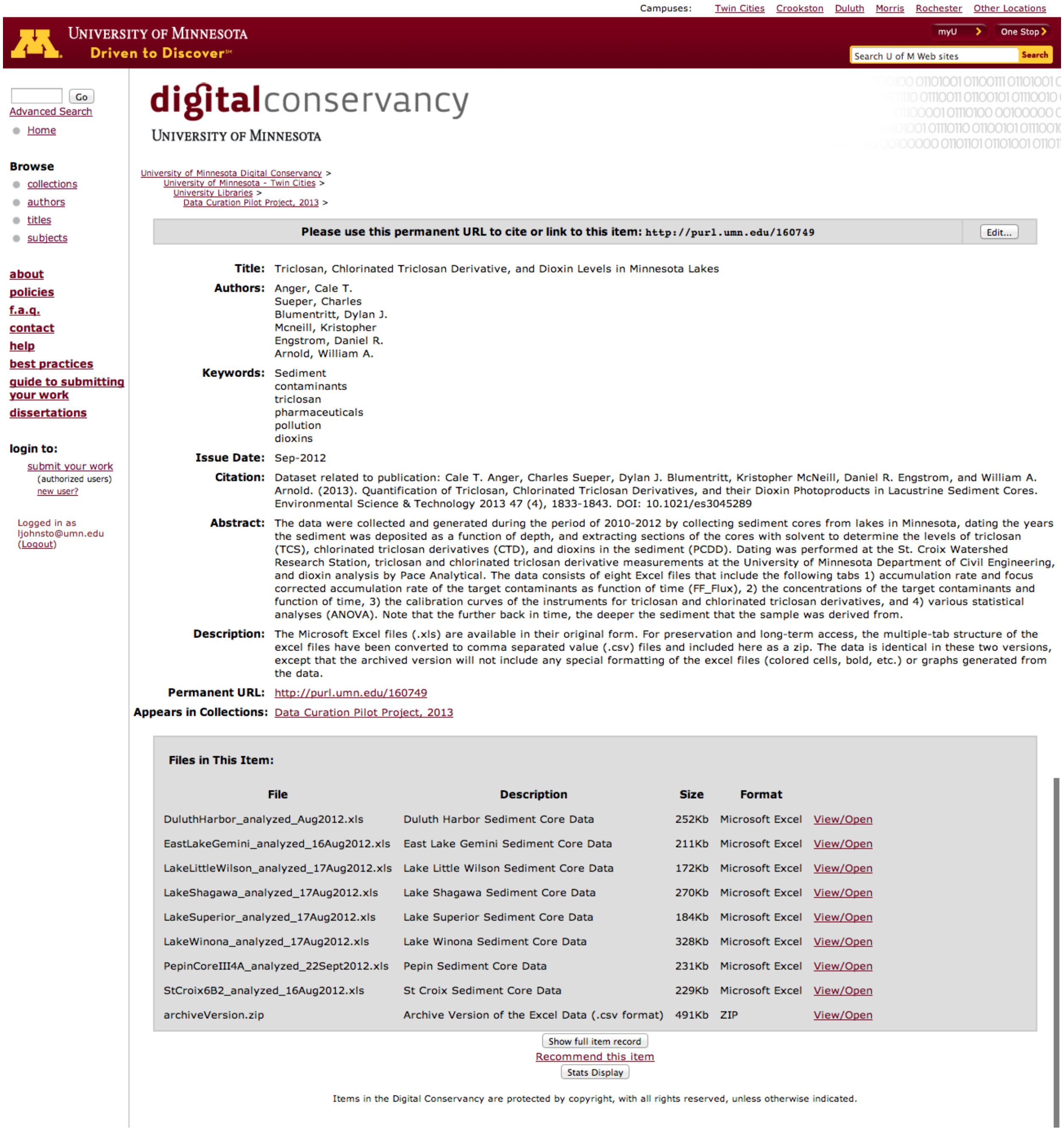
Interdisciplinary Data

This dataset (Figure 8) included Microsoft Excel data of chemical traces found in Minnesota lakes created by Bill Arnold (Civil Engineering, CSE) available at http://purl.umn.edu/160749. The original submission information (see appendix C.3) was curated in the following way:

Data Curation Stage	Actions Taken With This Dataset
0. Receive	● Data files received on 10-24-13. The deposit agreement was received (online) on 10-24-13. ● Author mentions additional "raw" data files that were proprietary and not included in the submission packet. The SIP was determined to be complete without the related files.
1. Appraise and Inventory	● Inventory high-level submission packet (8 Excel files, 1 thesis, author-submitted metadata) ● No related subject repositories for this dataset: UDC is appropriate home. ● Documentation provided in the form of a graduate student thesis, advised by the data author. Determined that the thesis could not be used as documentation of the dataset as the authors' consent was not given. Also, this MS thesis is not yet published in the UDC indicating that the author placed an embargo on the thesis – In the future, suggest including this in the SIP.
2. Organize	● Without additional documentation, the arrangement of the excel files would stand-alone. ● It may be possible to create a glossary of the headers used in the files, but this found to be out of scope for the pilot.
3. Treatment Actions	● Identify pros and cons of retaining current format. There was a lot of formatting and charts displayed in the excel files. However the data is of long-term value and a non-proprietary archival version should be captured. ● Each excel file (8) were opened and the tabs (ranging from 4-12 per file) were captured as .csv files to retain the data. These were all zipped into an archive file.
4. Description and Metadata	● Map author-submitted metadata to our metadata schema. ● Augment author supplied abstract with the common header terms used in the data. For example the code CTD was used for "chlorinated triclosan derivatives" and this was clarified in the abstract. ● Attach the description of the related publication in lieu of the dissertation, as this is not yet made publicly available.
5. Access	● Upload the 9 files to the UDC:

	○ Duluth Harbor Sediment Core Data (xls) ○ East Lake Gemini Sediment Core Data (xls) ○ Lake Little Wilson Sediment Core Data (xls) ○ Lake Shagawa Sediment Core Data (xls) ○ Lake Superior Sediment Core Data (xls) ○ Lake Winona Sediment Core Data (xls) ○ Pepin Sediment Core Data (xls) ○ St Croix Sediment Core Data (xls) ○ Archive Version of the Excel Data (.csv format) (zip) ● Include a description of the actions taken to create the archived file.
N. Reuse	● Notify the author of the dataset availability, the persistent URL location, and the recommended citation for reuse by others. ● Author is encouraged to track download statistics.

Figure 8: Screenshot of the resulting UDC record for the Interdisciplinary Data Example available at http://purl.umn.edu/160749.



Please use this permanent URL to cite or link to this item: `http://purl.umn.edu/160749`

Title: Triclosan, Chlorinated Triclosan Derivative, and Dioxin Levels in Minnesota Lakes

Authors: Anger, Cale T.
Sueper, Charles
Blumentritt, Dylan J.
Mcneill, Kristopher
Engstrom, Daniel R.
Arnold, William A.

Keywords: Sediment
contaminants
triclosan
pharmaceuticals
pollution
dioxins

Issue Date: Sep-2012

Citation: Dataset related to publication: Cale T. Anger, Charles Sueper, Dylan J. Blumentritt, Kristopher McNeill, Daniel R. Engstrom, and William A. Arnold. (2013). Quantification of Triclosan, Chlorinated Triclosan Derivatives, and their Dioxin Photoproducts in Lacustrine Sediment Cores. Environmental Science & Technology 2013 47 (4), 1833-1843. DOI: 10.1021/es3045289

Abstract: The data were collected and generated during the period of 2010-2012 by collecting sediment cores from lakes in Minnesota, dating the years the sediment was deposited as a function of depth, and extracting sections of the cores with solvent to determine the levels of triclosan (TCS), chlorinated triclosan derivatives (CTD), and dioxins in the sediment (PCDD). Dating was performed at the St. Croix Watershed Research Station, triclosan and chlorinated triclosan derivative measurements at the University of Minnesota Department of Civil Engineering, and dioxin analysis by Pace Analytical. The data consists of eight Excel files that include the following tabs 1) accumulation rate and focus corrected accumulation rate of the target contaminants as function of time (FF_Flux), 2) the concentrations of the target contaminants and function of time, 3) the calibration curves of the instruments for triclosan and chlorinated triclosan derivatives, and 4) various statistical analyses (ANOVA). Note that the further back in time, the deeper the sediment that the sample was derived from.

Description: The Microsoft Excel files (.xls) are available in their original form. For preservation and long-term access, the multiple-tab structure of the excel files have been converted to comma separated value (.csv) files and included here as a zip. The data is identical in these two versions, except that the archived version will not include any special formatting of the excel files (colored cells, bold, etc.) or graphs generated from the data.

Permanent URL: http://purl.umn.edu/160749

Appears in Collections: [Data Curation Pilot Project, 2013]()

Files in This Item:

File	Description	Size	Format	
DuluthHarbor_analyzed_Aug2012.xls	Duluth Harbor Sediment Core Data	252Kb	Microsoft Excel	View/Open
EastLakeGemini_analyzed_16Aug2012.xls	East Lake Gemini Sediment Core Data	211Kb	Microsoft Excel	View/Open
LakeLittleWilson_analyzed_17Aug2012.xls	Lake Little Wilson Sediment Core Data	172Kb	Microsoft Excel	View/Open
LakeShagawa_analyzed_17Aug2012.xls	Lake Shagawa Sediment Core Data	270Kb	Microsoft Excel	View/Open
LakeSuperior_analyzed_17Aug2012.xls	Lake Superior Sediment Core Data	184Kb	Microsoft Excel	View/Open
LakeWinona_analyzed_17Aug2012.xls	Lake Winona Sediment Core Data	328Kb	Microsoft Excel	View/Open
PepinCoreIII4A_analyzed_22Sept2012.xls	Pepin Sediment Core Data	231Kb	Microsoft Excel	View/Open
StCroix6B2_analyzed_16Aug2012.xls	St Croix Sediment Core Data	229Kb	Microsoft Excel	View/Open
archiveVersion.zip	Archive Version of the Excel Data (.csv format)	491Kb	ZIP	View/Open

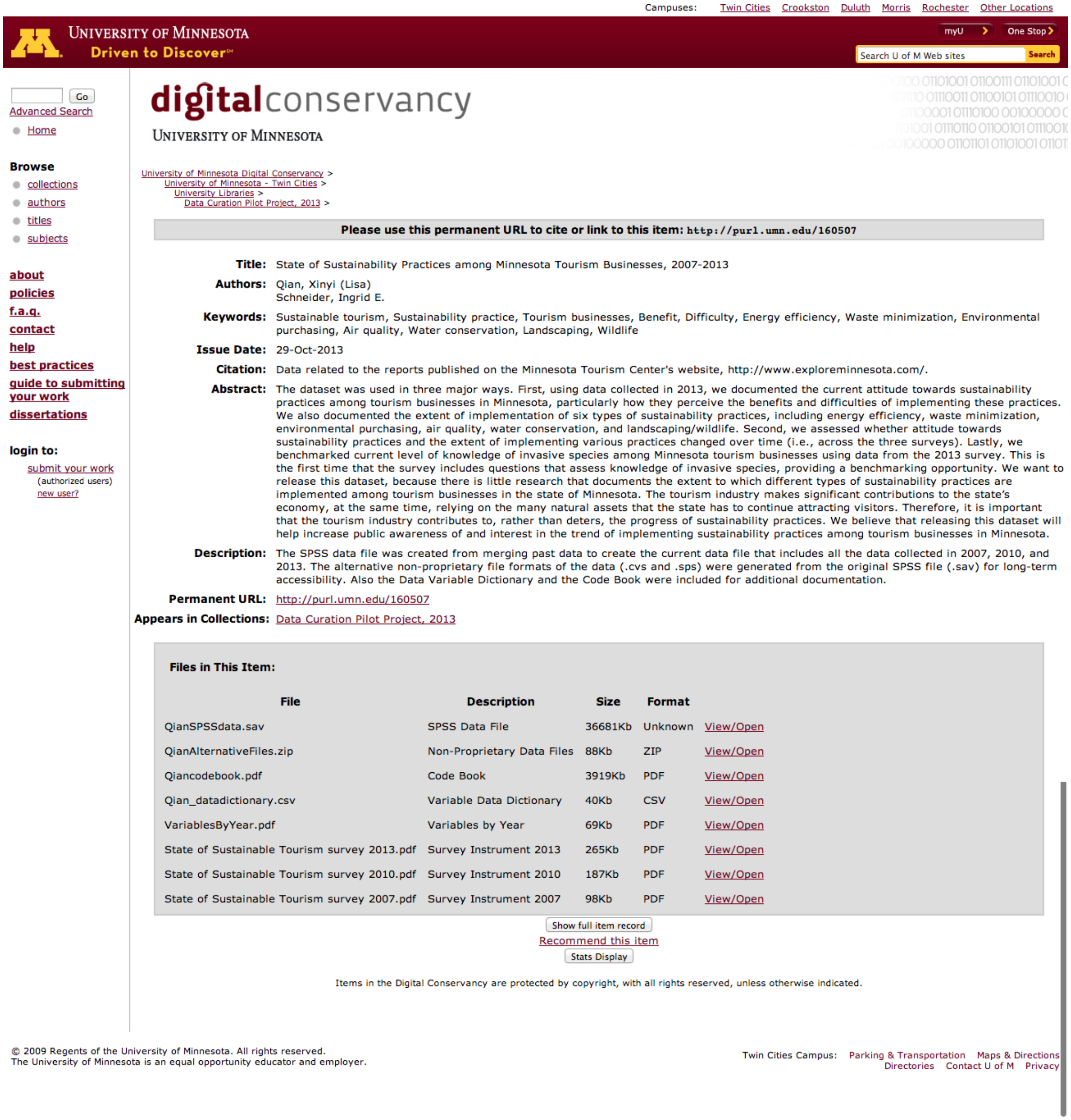
Natural Resources Data

This dataset (Figure 9) included SPSS data from online tourism surveys created by Lisa Qian (University of Minnesota Extension) available at http://purl.umn.edu/160507. The original submission information (see appendix C.4) was curated in the following way:

Data Curation Stage	Actions Taken With This Dataset
0. Receive	● Data files received on 10-25-13. The deposit agreement was received (online) on 10-29-13. ● Appropriate archive: It was determined that the appropriate home for this data would be the UDC.
1. Appraise and Inventory	● For the statistical files: ○ check for missing data/recodes/variant codes ○ are all the years identified? ○ do the questions change? ● The PDFs have some minor formatting issues. This is ok as is. ● The data dictionary is incomplete. This is actually only showing the variables used in each survey and if they were not included. not the variables themselves. ● Ask for survey "skip logic" documentation, if used.
2. Organize	● Create a working format.
3. Treatment Actions	● Preservation: ○ Convert initial data dictionary (excel) file to csv and rename to Variables by Year, to not confuse with the dictionary that will be generated. ○ Generate an SPSS syntax file and a .csv of the data that can be opened in non-proprietary software (such as R). The syntax file will read in the .csv and assign all the correct labels, values, and missing data codes to produce the intact .sav file. Zip these files.
4. Description and Metadata	● Create metadata for original SPSS format & non-proprietary format ○ Generate a data dictionary (the variable labels along with value labels) using SPSS syntax in a table by using the syntax "DISPLAY DICTIONARY." Save this as the Variable Data Dictionary. ○ Generate a codebook using the DISPLAY CODEBOOK syntax and saving as a PDF. ● Link to the published reports on the MN Tourism website. ● Map author-submitted metadata to our metadata schema.
5. Access	● Upload the 8 data files to the UDC: ○ SPSS Data File (.sas)

	○ Non-Proprietary Data Files (zip) ○ Code Book (PDF) ○ Variable Data Dictionary (PDF) ○ Variables by Year (PDF) ○ Survey Instrument 2013 ○ Survey Instrument 2010 ○ Survey Instrument 2007 ● Include a description of the actions taken to create the archived file in the record.
N. Reuse	● Notify the author of the dataset availability, the persistent URL location, and the recommended citation for reuse by others. ● Author is encouraged to track download statistics.

Figure 9: Screenshot of the resulting UDC record for the Natural Resources Data Example available at http://purl.umn.edu/160507.



Please use this permanent URL to cite or link to this item: `http://purl.umn.edu/160507`

Title: State of Sustainability Practices among Minnesota Tourism Businesses, 2007-2013

Authors: Qian, Xinyi (Lisa)
Schneider, Ingrid E.

Keywords: Sustainable tourism, Sustainability practice, Tourism businesses, Benefit, Difficulty, Energy efficiency, Waste minimization, Environmental purchasing, Air quality, Water conservation, Landscaping, Wildlife

Issue Date: 29-Oct-2013

Citation: Data related to the reports published on the Minnesota Tourism Center's website, http://www.exploreminnesota.com/.

Abstract: The dataset was used in three major ways. First, using data collected in 2013, we documented the current attitude towards sustainability practices among tourism businesses in Minnesota, particularly how they perceive the benefits and difficulties of implementing these practices. We also documented the extent of implementation of six types of sustainability practices, including energy efficiency, waste minimization, environmental purchasing, air quality, water conservation, and landscaping/wildlife. Second, we assessed whether attitude towards sustainability practices and the extent of implementing various practices changed over time (i.e., across the three surveys). Lastly, we benchmarked current level of knowledge of invasive species among Minnesota tourism businesses using data from the 2013 survey. This is the first time that the survey includes questions that assess knowledge of invasive species, providing a benchmarking opportunity. We want to release this dataset, because there is little research that documents the extent to which different types of sustainability practices are implemented among tourism businesses in the state of Minnesota. The tourism industry makes significant contributions to the state's economy, at the same time, relying on the many natural assets that the state has to continue attracting visitors. Therefore, it is important that the tourism industry contributes to, rather than deters, the progress of sustainability practices. We believe that releasing this dataset will help increase public awareness of and interest in the trend of implementing sustainability practices among tourism businesses in Minnesota.

Description: The SPSS data file was created from merging past data to create the current data file that includes all the data collected in 2007, 2010, and 2013. The alternative non-proprietary file formats of the data (.cvs and .sps) were generated from the original SPSS file (.sav) for long-term accessibility. Also the Data Variable Dictionary and the Code Book were included for additional documentation.

Permanent URL: http://purl.umn.edu/160507

Appears in Collections: Data Curation Pilot Project, 2013

Files in This Item:

File	Description	Size	Format	
QianSPSSdata.sav	SPSS Data File	36681Kb	Unknown	View/Open
QianAlternativeFiles.zip	Non-Proprietary Data Files	88Kb	ZIP	View/Open
Qiancodebook.pdf	Code Book	3919Kb	PDF	View/Open
Qian_datadictionary.csv	Variable Data Dictionary	40Kb	CSV	View/Open
VariablesByYear.pdf	Variables by Year	69Kb	PDF	View/Open
State of Sustainable Tourism survey 2013.pdf	Survey Instrument 2013	265Kb	PDF	View/Open
State of Sustainable Tourism survey 2010.pdf	Survey Instrument 2010	187Kb	PDF	View/Open
State of Sustainable Tourism survey 2007.pdf	Survey Instrument 2007	98Kb	PDF	View/Open

Twin Cities Campus: Parking & Transportation Maps & Directions
Directories Contact U of M Privacy

Social Sciences/Humanities Data

This dataset (Figure 10) included video and transcription files from Ojibwe conversations created by Mary Hermes (Curriculum and Instruction, CEHD) available at http://purl.umn.edu/160534. The original submission information (see appendix C.5) was curated in the following way:

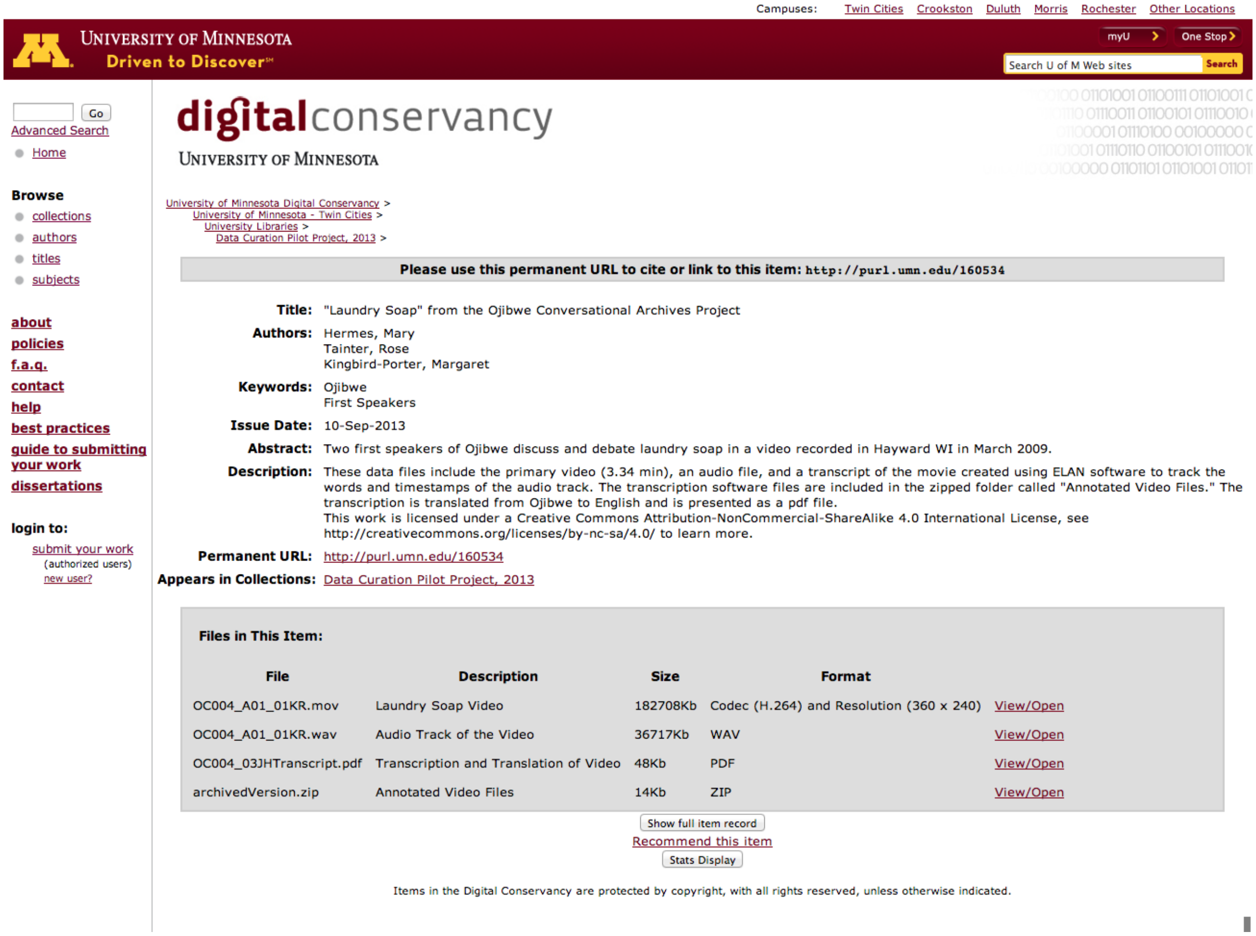
Data Curation Stage	**Actions Taken With This Dataset**
0. Receive	● Data files received on 10-28-13. The deposit agreement was received (online) on 11-27-13. ● The SIP did not originally include any author-supplied metadata fields. These were later supplied on 11-28-13. ● UMedia was considered for this dataset as it included a movie as the primary data file. The UDC was ultimately chosen to retain all pilot examples in one collection, but this should be revisited in the future.
1. Appraise and Inventory	● Verify file structure (need to download and use ELAN software) ● Identify the technical metadata of the AV files (audio/video codec) ● Verify research agreements (including IRB) of the individual participants in the video.
2. Organize	● Bring in copies of audio/video/ELAN files to test things ● How will we clarify "transcription" vs. "translation"?
3. Treatment Actions	● AV file treatment: ○ Evaluate if we need to transcode based on the standard vs proprietary codec used - min met. ○ Evaluate resolution to see if we need compress for web viewing/save space - keep as is. ● Preservation ○ The .mov and .wav file are in preferred formats for preservation. ○ Convert the MS word to PDF. ● Package software files (.eaf and .pfsx) with a readme file to describe how to run ELAN software. Save as zip.
4. Description and Metadata	● Export metadata from the ELAN file as an XML ● Create functional ReadMe.txt file of the preservation actions taken by the curators. ● Add resolution information to the video file. ● Use the description field to capture preservation actions taken and to include the author-recommended creative commons license.
5. Access	● Upload the 4 data files into the UDC: ○ Laundry Soap Video (mov)

  - ○ Audio Track of the Video (wav)
  - ○ Transcription and Translation of Video (PDF)
  - ○ Annotated Video Files (zip)
- The PDFs transcript and translation will be keyword searchable providing more comprehensive access to the data.
- Note the video and audio file must be downloaded in order to view, future in-browser viewers might assist this.

N. Reuse

- Notify the author of the dataset availability, the persistent URL location, and the recommended citation for reuse by others.
- Author is encouraged to track download statistics.

Figure 10: Screenshot of the resulting UDC record for the Humanities/Social Sciences Data Example available at http://purl.umn.edu/160534.



University of Minnesota Digital Conservancy >
University of Minnesota - Twin Cities >
University Libraries >
Data Curation Pilot Project, 2013 >

Please use this permanent URL to cite or link to this item: `http://purl.umn.edu/160534`

Title: "Laundry Soap" from the Ojibwe Conversational Archives Project

Authors: Hermes, Mary
Tainter, Rose
Kingbird-Porter, Margaret

Keywords: Ojibwe
First Speakers

Issue Date: 10-Sep-2013

Abstract: Two first speakers of Ojibwe discuss and debate laundry soap in a video recorded in Hayward WI in March 2009.

Description: These data files include the primary video (3.34 min), an audio file, and a transcript of the movie created using ELAN software to track the words and timestamps of the audio track. The transcription software files are included in the zipped folder called "Annotated Video Files." The transcription is translated from Ojibwe to English and is presented as a pdf file.
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, see http://creativecommons.org/licenses/by-nc-sa/4.0/ to learn more.

Permanent URL: http://purl.umn.edu/160534

Appears in Collections: Data Curation Pilot Project, 2013

Files in This Item:

File	Description	Size	Format	
OC004_A01_01KR.mov	Laundry Soap Video	182708Kb	Codec (H.264) and Resolution (360 x 240)	View/Open
OC004_A01_01KR.wav	Audio Track of the Video	36717Kb	WAV	View/Open
OC004_03JHTranscript.pdf	Transcription and Translation of Video	48Kb	PDF	View/Open
archivedVersion.zip	Annotated Video Files	14Kb	ZIP	View/Open

Items in the Digital Conservancy are protected by copyright, with all rights reserved, unless otherwise indicated.

Generalized Curation Workflow

Based on the experience of curating the pilot's example datasets, as well as the expertise and discussion shared at the Digital Curation Sandbox event, the project successfully resulted in a draft curation workflow of steps that might be taken by the library for archiving research data.

Stage 0. Receive

Our collection policies for research data should reflect our current services that include public access to digital objects. An version of our deposit agreement should be considered that incorporates adaptations for data.

	Activities for this stage	Questions to Consider
Step 1	**Receive the data**	
	Arrange for materials to arrive at library.	Are the files too big to obtain via email or Google Drive?
	Materials arrive at library.	Are all the necessary components of the data included for delivery?
Step 2	**Preliminary check**	
	Determine that the library is the appropriate repository for this type of data.	Is there a disciplinary repository for this type of data that should be deposited? If so, do we also accept a copy?
	Verify that data has no private or restricted information. Verify research agreements (with IRB + individual participants).	Do they meet our collection policies?
	Verify that the author has the necessary rights to deposit the data.	Is there any proprietary data or other copyrighted information?
	Understand any compliance issues (e.g., grant funder requirements) and verify expectations for reuse match our services.	What are the requirements for compliance (e.g., for NSF requirements). What are the requirements for reuse?
Step 3	**Receive any additional information**	
	Confirm or ask for a signed deposit agreement.	Do we need a different form for data?
	Collect any metadata about the data.	Do we use the same elements of the Dublin Core standard, or do we change the terminology depending on what type of data we are receiving?
	Collect E-mail correspondence related to data.	
	Create a submission information packet (SIP).	

Stage 1. Appraise and Inventory

	Activities for this stage	Questions to Consider
Step 1	**Secure the files**	
	Quarantine files (i.e., in a non-networked workstation).	Will we need to change these files? If so, do we need to create a working copy?
	Create duplicate backups if needed.	
	Perform any virus checks. Make sure we can actually open any files.	
Step 2	**Inventory the SIP**	Is this dataset complete? Are there future accretions?
	Create a manifest of submitted files	Are there any duplicate or unusable files?
	Identify the size of the data files.	Do we have the necessary storage space for these files?
	Identify file types.	
	Identify current organization, if any. Folder structure, file naming, etc.	
	Capture file creation date (modification dates will change over time with file transfers, etc.).	
	Verify file structure of complex files (GIS, HTML).	
	Note any limitations/alterations of the data file.	
Step 3	**Appraise the files**	Are there any versioning concerns/issues?
	Evaluate the data files for completeness and ability for reuse.	Is all the documentation there? Check for missing data/recodes/variant codes.
	Collect all known documentation related to material.	What documentation is provided? Is there information in published articles (e.g., methods section)?
	Check documentation for quality.	What things should be in the ReadMe file? Develop rubric?
	Identify any hidden documentation inherent to the file format.	Does the file format track all work on the file? Can we utilize it?
	Determine any files that do not need to be included with the data.	How do we tell good data from bad? Quality control?
	Verify all metadata provided by author.	
	Verify technical metadata (audio/video codec) that would be required for reuse.	

Stage 2. Organize

If elements were missing from the data (e.g., documentation of the data collection and preparation process), we asked for them at this stage.

	Activities at this Stage	Questions to consider
Step 1	**Select and/or request more information**	
	Determine library-based repository destination(s) for the data.	Should these be selected for the UDC, UMedia, USpatial, etc.
	Identify files with specialized software needs and determine if the software should be included with the SIP.	
	Ask for additional documentation from the creator if needed.	
	Remove any unnecessary files.	
Step 2	**Arrange and organize**	
	Determine which files are ready public consumption and which need further processing.	Which files are for administrative use?
	Identify relationships between files.	Which files are needed for creating metadata records?
	Create file structure (if none exists) that is meaningful.	Which files are the primary object and which are needed for reusing the dataset (supplementary?)?
	Discern file naming structure and describe if files should be renamed.	Do we keep file naming convention or rename?
	Determine the order of how files will be arranged for display.	How should the files be arranged for ease of access?

Stage 3. Treatment Actions / Processing

Note: It is important at this stage to create a record of treatment actions (i.e., add that information to preservation metadata).

	Activities at this Stage	Questions to consider
Step 1	**Consider file formats**	Does anything need to be moved to more preservation-friendly format?
	Determine what, if any, actions need to be taken regarding file formats.	Will we need to create derivatives of audio + video for faster streaming/download?
	Consider preservation policy when determining whether to retain original file formats or to normalize	Can the data be used by everyone in this format, if not, create an alternative access copy (eg. GIS to PDF)
	Identify pros and cons of retaining current format	Is there information that will be lost in transferring files from proprietary formats (e.g., pivot tables, formulas, color coding, etc.)? If so, keep two

		versions. For example, Excel files will with graphs and formulas will not convert these to .csv.. Need to capture this information in another way.
Step 2	**If necessary, convert to a preservation-friendly file format**	How do we mitigate risk in terms of reusability of these file formats?
	Follow established guidelines for preservation formats. • Create .xml (or other open source) versions for .docx or .eaf files. • Creating PDFs of word files (e.g. for making transcripts searchable). • Use open, nonproprietary formats if possible.	Do we keep both versions (submitted and normalized)?
	Consider accessibility issues. For example, caption and subtitle videos.	If you altered the format (e.g., for preservation), what would you lose?
Step 3	**Consider presentation**	
	Examine files for useable data labels -- create/change them, if needed.	Are there too many files for easy presentation or reuse? Can we package them in a way to convey their relationships?
	Rename files if necessary	Do we clean up the data?
	Clean the data if necessary	Is the data usable as it is presented?
Step 4	**Generate preservation metadata**	
	Create checksums for each file.	
	If necessary, preserve original dataset (offline and/or well-secured).	Do we have an archival copy and an access copy?

Ingest and Store

This stage involves the transfer of the data to an archive, repository, or other repository. In the case of the Data Curation Pilot, all datasets were ingested into our institutional repository, the University Digital Conservancy. Therefore, this step of the data curation workflow was handled as an aspect of the upload process, rather than staged activities for the data curator. In the absence of a repository solution, data should be stored in a secure manner adhering to relevant standards.

Stage 4. Description and Metadata

Apply author-generated metadata and create technical and provenance metadata.

	Activities at this Stage	Questions to consider
Step 1	**Create additional documentation/metadata**	
	Look at any metadata author included in collection.	Will we edit the supplied metadata?
	Determine disciplinary metadata schemas and vocabularies used in field, and if possible, present this metadata as a file separate from the dataset that might be used by information systems (e.g., XML).	Are there people involved in the project besides the researcher who could help with documentation/metadata?
	Create readme file with any actions a user would need to take to reuse the data.	Do we need to provide a template (common metadata framework) for documentation/metadata?
	Use tools such as DataUp to identify common/unique metadata elements in xls spreadsheets.	What will we do with the types the documentation files associated with the data?
Step 2	**Create repository metadata**	
	Create UDC record using Dublin Core metadata fields – single record with all individual files listed.	What are the gaps between author-supplied metadata and our metadata schema?
	Create short descriptions of each file for access and to define it among other files.	Will we add additional metadata?
	Consider how the files should be ordered since the UDC reverses the order they are entered for presentation (last one in, first displayed).	
Step 3	**Contextualize the data**	
	Create a collection name for the data.	What can we do to support links to published literature using the data?
	Create a collection-level description for the data.	What can we do to support citation of each editions?
	Contextualize (e.g., How does this fit with current data holdings?).	At what level of granularity will the collection be described?
	Identify related published works/identifiers.	How will we apply unique identifiers? To each creator? To each spreadsheet? To the whole dataset?
	Register dataset (DOI, etc.) – also identify related documents with DOIs, etc.	
	Apply identifiers to data (e.g., PURLs)	
	Identify/apply identifiers to creators (e.g., ORCID)	

Stage 5. Access

Ensure that data is accessible to both designated users and re-users, on a day-to-day basis.

	Activities at this Stage	Questions to consider
Step 1	**Publish to the web-accessible archive**	Do we need a rights-free version of the documentation for the dataset (e.g., if embedded in an article)?
	Complete deposit in repository systems (e.g., UDC).	
	Complete the upload of the data in the repository system, Remember that DSpace displays in the reverse order of upload. Plan accordingly.	
	If needed, get approval by administrator to "turn on" display.	
Step 2	**Enhance discoverability**	What kind of access will it have?
	Notify author of data availability, email link, etc.	Will there be an embargo? Is it open access?
	Full-text index the data (keywords) for searchability (e.g., .txt, PDF).	Any special requirements for access on part of funder? Repositories, etc.
	Expose metadata and keywords to search engines like Google and Google Scholar.	What if there's new editions/iterations of the data? Create a new record? Relate it to the original record?
	Broker metadata with databases such as MNCat, Worldcat, DataCite.	Do we support use of the data (questions about software, analysis techniques)?
	Push it to Experts@UMN or other UMN researcher profile tools.	
	Link it to the journal in which it was published.	
	Map to other collections (UDC).	
Step 3	**Track impact**	
	Track analytics of data, downloads, citations, etc.	
	Provide instructions for researchers to display stats and track their altmetrics.	
	Link to future publications/reuse examples.	
	Verify quality control through feedback from users and tracking use.	
Step 4	**Continue providing access for the long-term**	What support for ELAN software do we need/provide?
	Create long-term preservation plan.	Storing open source software to

		download from our site? Or emulation?
	Decide what formats we'll make the data available on request (if any).	Do we allow access to the hi-res image or video files? Charge at cost?
	Connect to versions of the data overtime.	How do we track the history of a dataset, versions? Relationships to other objects?
	Make necessary transformations/migrations of the data as needed.	

Stage N: Reuse

Data is reused or transformed. Track use metrics (e.g., number of downloads) and/or citations. Link to reuse examples from the data (websites, articles, transformations) that may provide additional context or detail.

Discussion of the Pilot's Successes and Shortcomings

Overall, the pilot was successful in its objective to identify, select, and pilot data curation services for five research datasets. This report and the resulting workflow successfully documented this process. In addition to five tangible examples of curated and accessible data, this project was successful in its more exploratory goals of piloting data curation as a service. Namely, the pilot successfully:

- Captured faculty expectations and needs through interviews and engagement activities.
- Involve the assistance of university staff (archivists, curators, digital technologists, metadata and cataloging staff, and subject-specific librarians) to share best practices for digital curation and establish a treatment process and workflow for curating research data in the library.
- Tested the current capacity of the University Libraries for curating research data by utilizing our existing infrastructure (e.g., the University Digital Conservancy).

The final objective of the pilot was to document the successes and shortcomings of our approach. Naturally, over the course of the pilot, issues surfaced regarding both our ability to curate the types of data that we received and the tools and infrastructure required to curate the data. For example, four out of the five datasets required specialized software (and the working knowledge to use the software) to open, investigate, and transform the files into preservation-friendly formats. Another concern was insufficient documentation included with the data. The researchers interviewed did not anticipate the level of documentation that would be needed to provide access to their datasets. Finally, data governance concerns, including ownership and intellectual property considerations, arose. Theses issues are each discussed before looking at the successes and shortcomings of our existing infrastructure used for curation, the UDC's DSpace-based repository tool.

Software Requirements and Preservation-Friendly File Formats

Nearly every dataset required the use of domain-specific software to convert file formats. Although these tools are commonplace in their respective research environments, they pose an interesting hurtle in the data curation process. For example, some are expensive and/or difficult to acquire, while others require specialized working knowledge to view and manipulate the files. The author sometimes relied on library and campus expert staff (who participated the Digital Curation Sandbox event) to open and work with these files. For example:

- The Natural Resources dataset required SPSS to open the .sas file. To supplement the data, the CLA Data Management specialist, Alicia Hofelich Mohr, used R (an open source statistical tool) to export the data dictionary and to create a non-proprietary version of the data file. This later process involved a custom-written script to export the data in a way that would be interoperable with non-proprietary statistical tools.
- The Engineering dataset included GIS files and required ESRI ArcGIS software to open and manipulate. Borchert Map Library director, Ryan Mattke, did the bulk of the work to identify the correct files, convert them to shape files (which required a specialized add-on to the software), and, with his specialized skills in map creation, to generate a view of the final map that is functional for users without GIS tools. Mattke also identified missing metadata (definitions of the numeric codes used in the map). After contacting the relevant researcher, we were able to augment the metadata to include this important element.

On the other hand, several datasets required software that is more readily available to curation staff, yet presented some unanticipated challenges. For example:

- ELAN is an open source software tool that was used to create the transcription files in the Humanities and Social Sciences dataset. This tool was successfully downloaded from the Language Archive (http://tla.mpi.nl/tools/tla-tools/elan/) and used in the curation process. However, although it is currently supported by the grant-funded project, ELAN is not well known outside of the field, and the future of the tool's availability is difficult to predict. Therefore, the ELAN file formats were included alongside XML-based versions of the files to better ensure long-term access.
- The Health Sciences dataset files were generally straightforward, with the bulk of the conversion into preservation friendly formats happening between commonly available Microsoft products. However, the Manual of Practices MS Word file proved to have additional complexity when saved as a PDF. The document was written with section splits (a Microsoft Word feature), which resulted in multiple PDFs when it was imported to that format. The many PDFs were then combined into one using the Adobe Acrobat Pro software, a tool not available on most staff workstations in the Libraries.

As future data curation services are developed, it will be important to consider the variety of software – and expertise to use the software – required for the wide range of disciplines at UMN. This pilot demonstrates that multiple people with various expertise were required to do the necessary data curation work. Data curation services will require Libraries staff to be knowledgeable and proficient in the commonly used software tools; those especially important in this study were statistical tools (SPSS, R) and GIS software (ArcGIS).

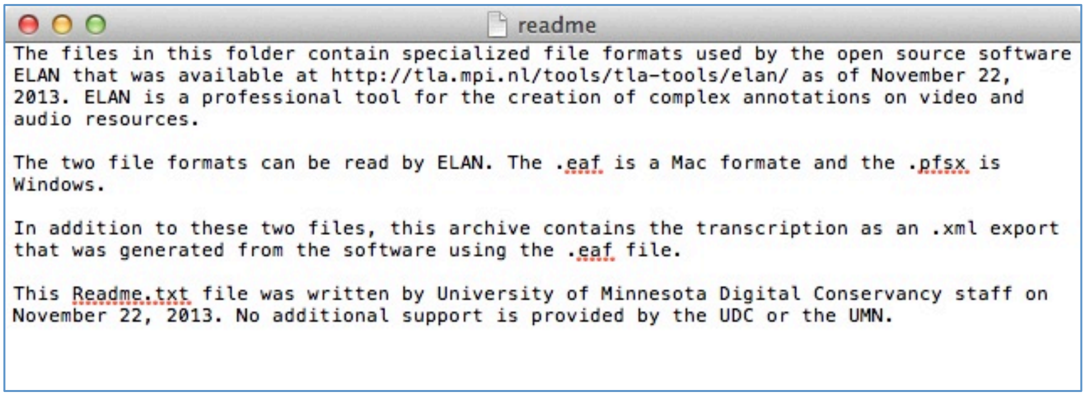
Documentation Quality and Metadata Limitations

Next, the quality and level of documentation for research data varies widely among disciplinary and individual data management practices. Anecdotally, researchers find that a primary barrier to releasing research data for access is the lack of information (metadata, documentation, etc.) that would support future reuse. In other cases, the lack of documentation may not be a concern, but a protection, so that only those who understand the data will be able to use it. In either case, the Libraries must support researchers in the creation of documentation and metadata that supports their goals for their research data, and if accessioned into the Libraries' collection, documentation that also supports broader reuse of that data.

An excellent example of quality documentation was included with the Health Sciences dataset. The 237-page Manual of Procedures includes not only the study design and data collection instruments, but also contextualizes the research with detailed procedures, data management techniques, and a review of the polices and reports associated with the multi-year study. In addition to this complete write-up of the data collection process, a data dictionary describing the unique codes used in the data files supplements the dataset. In one example of its utility, the data dictionary describes the code PID as "Participant ID (first digit indicates center; next 4 digits are sequential; last digit is a check digit)."

Documentation also came in more unconventional forms, such as references to related (traditional) publications. Four of the datasets included links to published articles; this information is displayed in the repository metadata record. It may be possible for data curators to capture excerpts from those publications (e.g., the "methods" section of a research article) to help provide the needed context for the data held in our repositories. Another unique descriptive file was generated for the GIS dataset, a PDF view of how the GIS data might look as a map. This context will allow a user to determine, geospatially and visually, if the dataset might be useful to them. Finally, supplemental documentation was written by curation staff and included with the dataset as a readme.txt file that explains how the dataset file formats were arranged for reuse and which tools can be used to use and transform it (see Figure 11).

Figure 11. A documentation file generated by library staff for the 2013 Data Curation Pilot.



The files in this folder contain specialized file formats used by the open source software ELAN that was available at http://tla.mpi.nl/tools/tla-tools/elan/ as of November 22, 2013. ELAN is a professional tool for the creation of complex annotations on video and audio resources.

The two file formats can be read by ELAN. The .eaf is a Mac formate and the .pfsx is Windows.

In addition to these two files, this archive contains the transcription as an .xml export that was generated from the software using the .eaf file.

This Readme.txt file was written by University of Minnesota Digital Conservancy staff on November 22, 2013. No additional support is provided by the UDC or the UMN.

Metadata limitations were also felt as a shortcoming with the Dublin Core (DC) schema of the libraries' repository system. However, to supplement the minimal description provided by DC, in two cases, the domain-specific metadata of the dataset were copied to a separate file to provide human-readable context as well as exposing this information to the full-text indexers, should users be searching for a particular term in the schema. This took place for the Engineering GIS dataset, with FGDC-compliant metadata embedded in the GIS files and the Humanities/Social Sciences transcription files, which contained XML-based metadata generated by the transcription software ELAN. In both cases, curators were required to export these metadata assets and save them as external plain-text files.

On the other hand, in several cases, the description and abstract fields of the DC-based metadata of the repository software were the only form of descriptive documentation. For example, the Interdisciplinary Sciences dataset included documentation in the form of a graduate student thesis. This item could not be included due to ownership concerns (more details in the data governance discussion are below), however, much of the data's context was missing from the primary excel files. DataUp (http://dataup.cdlib.org), an open source Microsoft Excel add-on for creating metadata from tabular data, was considered for capturing information about the dozen .xls files in this dataset. Unfortunately, several attempts at logging in and testing this tool were unsuccessful since the web-based DataUp system was either down or not working in November 2013. The add-on version is currently only available for PCs. Therefore, the abstract for the record was augmented to define any acronyms used in the data, and a description field was generated to document the curation steps taken by library staff to transform the files into non-proprietary formats.

When implementing data curation services, the Libraries should prepare to use a variety of metadata schema and documentation templates to assist researchers in creating quality documentation. More detailed metadata schema that would have augmented the datasets in the pilot, such as extensible DDI[14] or reputable domain-specific schema such as ecology-based EML.[15] In addition, metadata templates or web-forms that are outside of the repository may allow for further customization of the field required for author data submission and might be discipline specific.

Data Governance Concerns

The governance of research data, involving how data can be made available and with whom, is of great concern to researchers. Issues such as private data, limitation on use, and questions of data ownership each came into play just within our small sample of datasets.

Although private data was explicitly excluded from our pilot, the fact remains that not all data should be released openly to the public, and the various level of access are not yet well understood by our researchers. In the curation workflow model, it is good practice to check and verify that no private data exists when ingesting new files. This is not an easy task. Simple scripts might be used to check for obvious problems (e.g., birth dates, social security numbers, etc.), but these assume the data incorporate this information with correct formatting. Also, the submission agreement for authors contributing their digital works to be curated in any library generally includes a clause that the information does not contain private data. However, libraries curating research data may inadvertently accept private data, and must weigh this risk at all times.

Related to this topic of private data, researchers also may need deidentification services or consultation. This necessary step should happen before the curation process, and data curators should work with other service providers on campus to better support researchers in all disciplines that deal with human subjects. For example, three of our datasets involved human participants. How the release agreements are written, what level of consent participants give, and any communication with the institutional review board ideally should be better identified and recorded during the curation process. Although this step was discussed with the data authors, in the pilot there was no documentation captured to indicate the level of consent participants agreed to for data dissemination. During the Digital Curation Sandbox, it was noted that the Libraries' Gift Acceptance forms (for the archives) and the digital repository's deposit agreements should be rewritten to add greater weight to this issue. An example might be taken from Harvard's web-based data repository, DataVerse.[16] This tool has an interesting use clause: in order to download data from the repository, users must agree not to use the data to re-identify its subjects in any way – thus acknowledging the limitations of standard deidentification techniques.

[14] The Data Documentation Initiative (DDI) metadata profile built for the social and behavioral sciences can be downloaded for use at http://www.ddialliance.org/.

[15] The Ecology Metadata Language (EML) is a detailed XLM-based schema that is adaptable for many physical and biological sciences. See http://knb.ecoinformatics.org/software/eml.

[16] The Harvard Data Verse Network is open to public access at http://thedata.harvard.edu/dvn/ and the Terms of Use appear when downloading a file.

Next, data governance issues do not only stem from restrictions, but also from access requirements. As noted in the introduction of this report, the 2013 Data Curation Pilot Project is, in part, testing out the Libraries' current capacity to support upcoming federal requirements for researchers to make the resulting data from federal grants (e.g., NSF, NIH) more publically accessible. The pilot interview with faculty included questions related to what, if any, expectations the funders had for data sharing. In fact, one notable response by a faculty member to our pilot's initial call for proposals was, "This is a good idea. Proposals now have requirements for this." However, as the federal requirements evolve, it is important to consider how they match the interests of all parties involved: the data authors, the repository, and the institution. In some cases, there may be limitations on distribution that the repository is not able (or willing) to make. For example, out of the 16 responses to our call for proposals in the pilot, 3 responded that they included private or restricted information (i.e., data that was identified in such a way as to prevent openly public distribution). Additionally, 38% (6 out of our 16 respondents) indicated with their interest to participate in the data curation pilot that they "do not necessarily need public access to [their] data." Conversations about data openness and stakeholder expectations are an important element of data curation.

For example, the Social Sciences/Humanities researcher, after seeing the resulting dataset curated for use in the repository, voiced concerns about the wide availability not only of access, but also of use. The data curator explained the options, which included Creative Commons licenses[17] that encourage reuse, but only by those who would not use their data for commercial gain. A suitable license was chosen by the researcher and incorporated into the dataset record. This desire for more flexibility over data access and use will continue as more researchers are faced with increased requirements to make their data publically available.

Intellectual property of research data was another concern of the researchers in our pilot. Several questions arose in this area early in the pilot during our faculty interviews. For example, the author of the Natural Resources data was uncertain if the instrument used in her surveys could be included in the data release. Though datasets may not be subject to copyright under U.S. law, UMN researchers are able to license their survey tools and other commercializable products. In this case, it was found that the survey tool could be published. Future curation services should consider whether data assets and their related products (e.g., software, educational materials, videos, etc.) are or will become intellectual property whose value may be affected by becoming freely available via the web.

Finally, it is very interesting to note that in at least two instances within our five datasets, the project encountered objects that were not authored or owned by the faculty. The Engineering dataset included a scanned map that was authored by a

[17] Creative Commons licenses describe to others how you allow your work to be used. See http://creativecommons.org

state agency in the 1950s and may still be under copyright. Data that was not authored by the contributor or a UMN author brings a the repository a potential risk of copyright violation for public distribution on the web. In this case, the researcher confirmed that the state had an interest in making this data public, and the library staff determined that the work was transformative, thereby satisfying requirements for fair use. Library staff therefore uploaded the map. However, even in a low risk situation as this, data curation services must communicate to authors the importance of data ownership issues.

The second example presented a greater risk. The documentation for the Interdisciplinary dataset included a digital copy of a recent master's thesis by a graduate student who worked extensively with the data. This thesis is useful because it provides a wealth of context for how the data was collected, analyzed, and used. In fact, the Libraries should consider a service that links deposited datasets to any resulting theses, as all UMN theses are digitally deposited into the institutional repository. However, the consent of the student is mandatory. In this case, the student was not consulted in the data pilot by the faculty author, and the curation staff found that this particular thesis appeared to be under embargo (at the student's request). Therefore, the thesis is not linked to the dataset at the present time.

DSpace as a Tool for Data Curation

The UDC currently runs on DSpace software (http://www.dspace.org), first implemented at UMN in 2007 to manage the University's scholarship, mainly publications, for public access. As a tool for data curation, however, this tool is extensible, with a few minor user-interface flaws (most notably the difficulty of reordering files once they are uploaded to a record, an issue that is resolved with the newest version of DSpace, which was implemented after the pilot's completion in December 2013). However, two shortcomings were noted arising from the DC metadata profile of the software.

First was the challenge of accurately displaying and accommodating several files that relate together in a complex way, within a single data record. For example, each dataset included some variation of the data that was captured for preservation purposes (e.g., a CSV version of an Excel file). These archival versions of the data were uploaded with the original data, but their purpose was not immediately clear unless each was described with a term such as "Archival version of the dataset." Also, with multiple files to display, the archival version could quickly overwhelm the primary data files, as was the case with the Interdisciplinary dataset. To circumvent this problem using DSpace, the archival files were zipped together to form one downloadable file. This is likely not the best solution for long-term preservation, as the formats of individual files within a composite zip file becomes obscured, as well as the potential for data degeneration due to lossy compression. However, for the purpose of arrangement and display, zipping distinguished each data file and allowed users to choose which to download.

Second is the need to tie the data record to the publications that analyze and report its resultant research. For example, the Health Sciences dataset included a file-listing article published using the data. Any articles that accumulate in the future should also be tied to the dataset record in some way. Tracking these relationships between data and published research is not an easy task. The workflow is not designed to support giving permissions to researchers in order for them to edit their dataset record with new publication information. A more dynamic workflow should be explored that would allow for future uses of the data to cite the persistent URL and allow for trackbacks to appear in the record. On the researcher side of things, disambiguation registries such as ORCID[18] might be employed to track future, and potentially related, publications and datasets.

Overall, DSpace was able to successfully accommodate the variety of preservation actions that were recommend in the general workflow. These included: generating checksums, verifying file extensions, providing human-readable descriptions, and capturing metadata in a standard form. An additional feature that may be of use in the future is incorporating a persistent identifier for data citations. For example, DSpace 4.0 supports DataCite,[19] which provides DOIs that may assist researchers with integrating datasets into their published scholarship. These links can demonstrate the impact of research within a field for tenure and review purposes.

One consideration in using DSpace in the data curation process is how and when metadata is captured from the researcher. Files that are received are not necessarily the final, processed files that are uploaded to the repository. In this pilot, the metadata was provided with the dataset in the Receiving Stage, long before the data was ready for upload. Therefore, a data curation tool incorporating DSpace, which is built for self-deposit without intermediary curation, must bridge the gap between user-submission metadata and curation staff processing actions, without breaking the workflow. For example, a web-based form or a SWORD-based interface might be used to capture the metadata apart from the file ingest.

Conclusions of the 2013 Data Curation Pilot Project

This project set out to curate example datasets of UMN research data in order to better understand the needs of researchers and other data users, the effort involved with curation, and the process or workflows for curating data, thereby testing our existing capacity. In each of these elements, the pilot was a success, and we now have a solid base of knowledge from which to build future data curation initiatives.

[18] ORCID provides researchers with a persistent identifier to support linkages between their works more consistently than names and affiliations. See http://orcid.org/

[19] DSpace 4.0 Documentation details the use of DOIs in the latest release at https://wiki.duraspace.org/display/DSDOC4x/DOI+Digital+Object+Identifier

Following the work of this pilot, the real challenges begin. Future activities include engaging the participation of data users and potential partners, including the Office of Information Technology (OIT) and the Office of the Vice President for Research (OVPR), to investigate implementation of data services. Next steps might include:

- Identifying and clarifying the various roles of the campus entities as they relate to data curation activities, such as data governance issues, preservation functions, and domain knowledge to support use of curated datasets.
- Recommending infrastructure to support data curation services beyond our current capacities.
- Identifying potential ways to embed data curation workflows into existing research life cycles in order to respond to data sharing mandates by UMN or federal agencies.
- Developing Libraries staff to support data curation activities that are illustrated in the workflow model, particularly expertise in research software.
- Researching, testing, and developing a cost model to support the ongoing expenses of curating research data.

This report summarizes the steps taken to curate the five datasets in our pilot. As a result of this project, the University Libraries are better prepared to fine-tune and implement selected recommendations from previous assessments and committee reports. Additionally, the Libraries now hold a more realistic sense of the overall capacities needed to develop a sustainable data curation service model. Due to variables of scale, domain-specific data requirements, and diversity of domain culture and practices, the success of such a model will likely depend upon strong collaboration among interdependent service providers. To be successful, significant capacities in the areas of data management and curation, infrastructure, and domain knowledge must coalesce in operationally effective ways that minimize barriers to and demands on researchers. All this in the hope for increased public access to federally funded research.

Appendices

APPENDIX A: DATA CURATION PILOT PROPOSAL

Data Curation Project Proposal

2013 President's' Excellence in Leadership Project Participant
Lisa Johnston

Summary

The Data Curation Project will expand the University Libraries programmatic efforts to support research data management requirements by establishing a fixed-term data curation pilot with the objective of developing archiving workflows for a variety (3-5) of types of research data. This pilot will utilize our current suite of services and expertise in the Libraries. Specifically, in eight months, this project will result in 1) a data curation workflow for archiving research data in the University Digital Conservancy and 2) and summary report describing the successes and shortcomings of this approach.

Statement of Need

The University Libraries offers data management services to the campus including consultation on writing data management plans, and tools and solutions for creation, storage, analysis, dissemination, and preservation of research data. Our services were developed in 2010 in response to expressed faculty need and a data sharing mandate from the National Science Foundation that requires data management plans to accompany all grant proposals. Recently (Feb 2013), the federal Office of Science and Technology Policy issued a memorandum directing Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research. In February 2013, the Fair Access to Science and Technology Research Act (FASTR) was introduced in both the U.S. House of Representatives and the U.S. Senate. This bi-cameral and bipartisan legislation would require federal agencies with annual extramural research budgets of $100 million or more to provide the public with online access to research manuscripts stemming from funded research no later than six months after publication in a peer-reviewed journal.

The Libraries manage the University Digital Conservancy, which provides free, worldwide open access to scholarly and administrative works produced by or about the University of Minnesota. We anticipate increased campus needs for data management and repository services, given current directions of the federal government. This project will utilize our current suite of services and expertise to expand our programmatic efforts to accommodate a variety of needs from scholars across the disciplines to pilot our data curation capacities and determine a set of potential workflows the University Libraries might adopt and expand. Curation includes appraisal, ingest, arrangement and description, metadata creation, format transformation, dissemination and access, archiving, and preservation.

Amid all of these known needs and potential roles for the University Libraries, we must remember that the University community (including library staff) may not immediately see the library as a primary service provider in this area of data management and curation. Throughout, this project will also implement change management techniques to help engage stakeholders (faculty, researchers, UL staff, and students) toward a shared view that the library is a partner in the research data management process.

Objectives and Success Criteria

Objectives of project:

1. Pilot a data curation workflow for 3-5 research data examples.
    a. Identify technical requirements for pilots (eg. files smaller than 1GB, a variety of disciplines, can be made publically available); select data that meet these criteria to help guarantee successful workflows.
    b. Involve the assistance of library liaisons to partner with and establish relationships with the faculty data owners.
    c. Explore and document faculty expectations and needs (Eg. what services do they need, how does cost factor in?).
    d. Curate the 3-5 pilot projects using existing infrastructure (eg. UDC) and document this process.
2. Write a report on the experience of the project including the successes of the approach and the shortcomings.

Success Criteria: As a result of this project:

1. The University Libraries are better prepared to make recommendations and implement sustainable data curation services.
2. A shift in perceptions about what services the library can provide to support research at the UMN.

Scope and Out-of-Scope

In scope:

- The project is open to exploring one or more potential research data curation workflows using the existing infrastructure or potential openly available tools (eg. DMPTool, DataUp).
- All disciplines should be considered for the 3-5 pilots and be chosen based on:
    - availability of research data files to curate within the project timeline;
    - willingness of the faculty owner to work with us in a “pilot” atmosphere; and
    - best match to our current policies and requirements for archiving.
- If pilots are not found for previously unavailable data, it is possible to “reload” the research data already held in the UDC to document this process in a systematic and thoughtful way.
- Exploring faculty and staff attitudes on potential data curation services through interviews and discussion events.
- Partnering with new staff in Data management roles (eg. DAH CLIR fellow; DM specialist for CLA-OIT).

- Exploring alternative approaches to data curation underway at peer-institutions.

Out-of-scope (These would be new projects and need participation by more stakeholders):

- Recommending infrastructure to support data curation services beyond our current capacities.
- Identifying potential workflows for curating research *publications* in response to the possible OA mandates by the university or the federal agencies.
- Developing a cost-model for curating research data.

Schedule (High Level with communications plan)

Project Timeline: May - December 2013.

Timeline (2013)	Project Milestone	Communications
May	Draft **work plan**; Establish pilot **criteria** for the type of data/relationship.	Share **Proposal, Work plan, and criteria** with Sponsor; Announce project to UL (MM)
June	Data Curation **Event** for UL staff; Identify pilot data opportunities.	Email Pilot invitations to potential faculty
July	Pre Curation **Interviews** with faculty (if possible)	
August	Research data curation programs and workflows (at peer institutions); lit review	
September	Develop **Workflows** and **Archive** data in UDC.	
October	Post Curation **Interviews** with faculty.	
November	Draft **Report** on outcomes of project.	Share Draft **Report** to UL Cabinet
December	**Present** Project at PEL Closing Reception (Dec 3, 2013)	Share Final **Report** to UL staff (MM), Archive in UDC

Stakeholders/Sponsors

Project Lead: Lisa Johnston (as part of the PEL 2013 project)

PEL Circle Mentor: Brian Herman (July-December monthly meetings)

Project Co-Sponsors: Karen Williams and John Butler

Stakeholders: Library Administration (Cabinet); UDC co-director; Liaisons to the pilot data owners; data preservation specialist; metadata strategist; copyright librarian; R&L directors.

Communications Audiences: Library staff; partner service units (OVPR, OIT, CLA-OIT, MSI); research faculty and staff.

Outcomes/Deliverables

The minimum products of this project are:

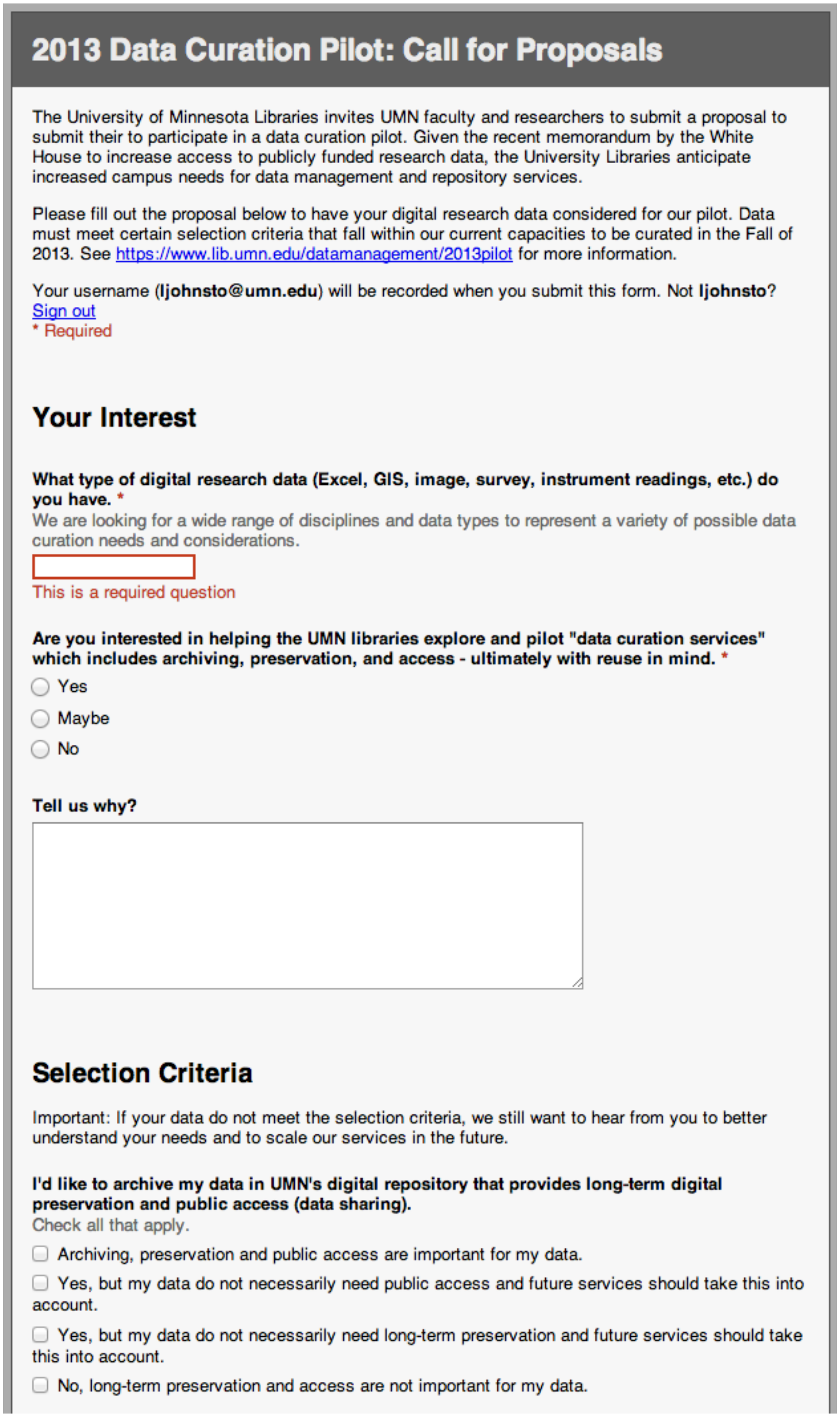
- Project Proposal (Due May 16, 2013)
- Project Work plan
- Call for proposals statement (with selection criteria) (for review by May 30th, 2013)
- Documentation on possible workflows for Data Curation (internal)
- Final report on the lessons learned
- Presentation on the project outcomes (due Dec 3, 2013 at the PEL Reception)

References and Background

The emerging service of Data Curation was discussed as a component of the Research Services Framework: written by a subgroup of the R&L Directors (Jan 3, 2013). Karen Williams called a broader meeting to discuss these emerging service areas in more detail (Feb 8, Feb 26, 2013). Several library reports were reviewed and used for this the Data Curation portion of the framework; they include:

> 2012: Agenda for Deepening Library Support for Research by John Butler, Layne Johnson, and Lisa Johnston for the ARL E-science Institute, 2011-12 Cohort.
>
> 2011 Near-Term Recommendations for Action from the Data Management, Access, and Archiving Working Group (Subgroup of the Libraries RSSC)
>
> 2009 Data Stewardship Opportunities for the University of Minnesota Libraries Recommendations from the E-Science Data Services Collaborative (of EDSC)

APPENDIX B: CALL FOR PROPOSALS FORM INSTRUMENT



2013 Data Curation Pilot: Call for Proposals

The University of Minnesota Libraries invites UMN faculty and researchers to submit a proposal to submit their to participate in a data curation pilot. Given the recent memorandum by the White House to increase access to publicly funded research data, the University Libraries anticipate increased campus needs for data management and repository services.

Please fill out the proposal below to have your digital research data considered for our pilot. Data must meet certain selection criteria that fall within our current capacities to be curated in the Fall of 2013. See https://www.lib.umn.edu/datamanagement/2013pilot for more information.

Your username (**ljohnsto@umn.edu**) will be recorded when you submit this form. Not **ljohnsto**?
Sign out
* Required

Your Interest

What type of digital research data (Excel, GIS, image, survey, instrument readings, etc.) do you have. *
We are looking for a wide range of disciplines and data types to represent a variety of possible data curation needs and considerations.

This is a required question

Are you interested in helping the UMN libraries explore and pilot "data curation services" which includes archiving, preservation, and access - ultimately with reuse in mind. *

○ Yes

○ Maybe

○ No

Tell us why?

Selection Criteria

Important: If your data do not meet the selection criteria, we still want to hear from you to better understand your needs and to scale our services in the future.

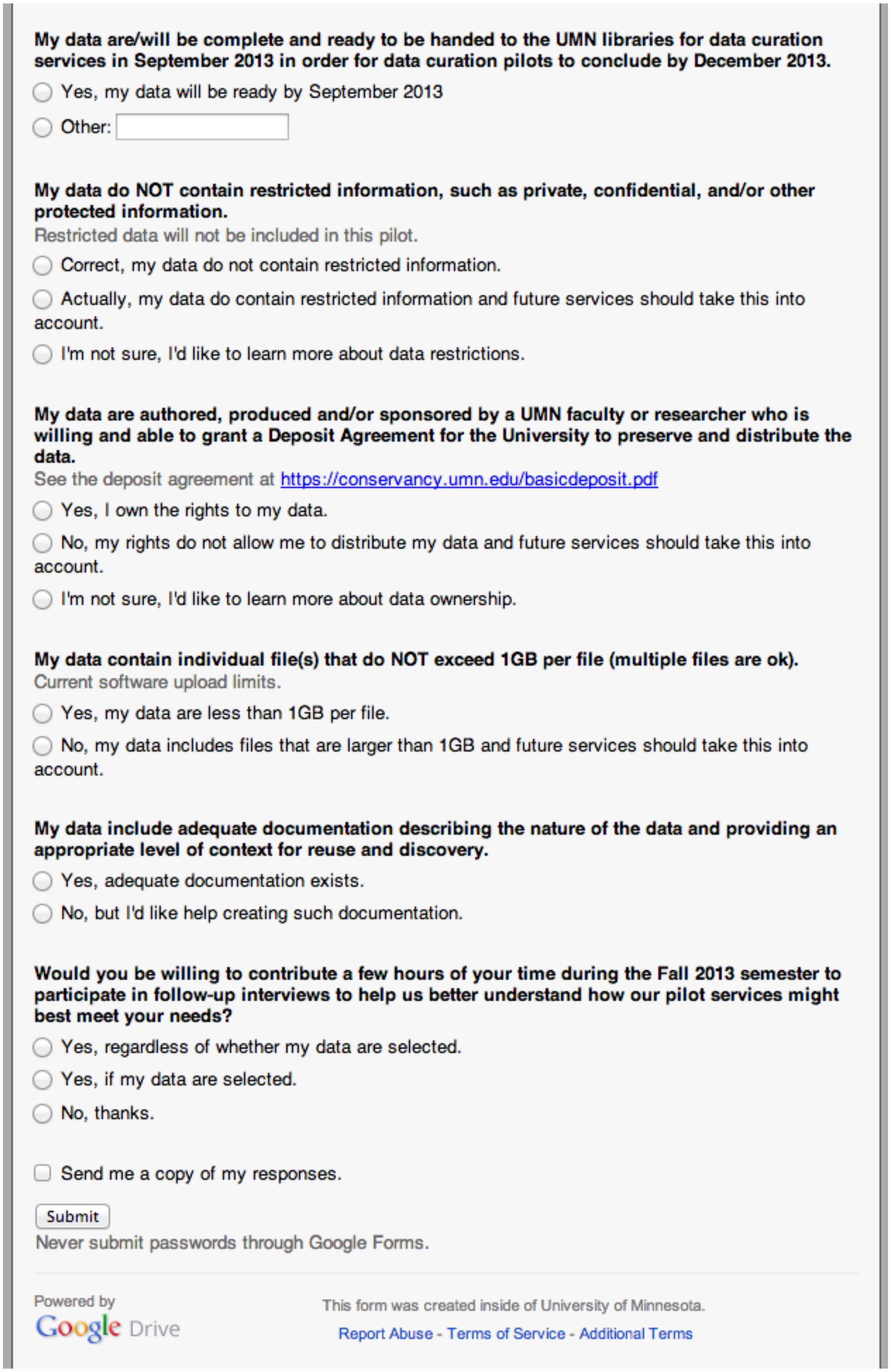
I'd like to archive my data in UMN's digital repository that provides long-term digital preservation and public access (data sharing).
Check all that apply.

□ Archiving, preservation and public access are important for my data.

□ Yes, but my data do not necessarily need public access and future services should take this into account.

□ Yes, but my data do not necessarily need long-term preservation and future services should take this into account.

□ No, long-term preservation and access are not important for my data.



My data are/will be complete and ready to be handed to the UMN libraries for data curation services in September 2013 in order for data curation pilots to conclude by December 2013.

○ Yes, my data will be ready by September 2013

○ Other: ______

My data do NOT contain restricted information, such as private, confidential, and/or other protected information.

Restricted data will not be included in this pilot.

○ Correct, my data do not contain restricted information.

○ Actually, my data do contain restricted information and future services should take this into account.

○ I'm not sure, I'd like to learn more about data restrictions.

My data are authored, produced and/or sponsored by a UMN faculty or researcher who is willing and able to grant a Deposit Agreement for the University to preserve and distribute the data.

See the deposit agreement at https://conservancy.umn.edu/basicdeposit.pdf

○ Yes, I own the rights to my data.

○ No, my rights do not allow me to distribute my data and future services should take this into account.

○ I'm not sure, I'd like to learn more about data ownership.

My data contain individual file(s) that do NOT exceed 1GB per file (multiple files are ok).

Current software upload limits.

○ Yes, my data are less than 1GB per file.

○ No, my data includes files that are larger than 1GB and future services should take this into account.

My data include adequate documentation describing the nature of the data and providing an appropriate level of context for reuse and discovery.

○ Yes, adequate documentation exists.

○ No, but I'd like help creating such documentation.

Would you be willing to contribute a few hours of your time during the Fall 2013 semester to participate in follow-up interviews to help us better understand how our pilot services might best meet your needs?

○ Yes, regardless of whether my data are selected.

○ Yes, if my data are selected.

○ No, thanks.

□ Send me a copy of my responses.

Never submit passwords through Google Forms.

This form was created inside of University of Minnesota.

Appendix C: Pilot Data Submission Information

Based on the Call for Proposals responses and the pre-curation interviews with each data author, these submission information packets (SIPs) were created and distributed as summary handouts to each team in the Digital Curation Sandbox Event.

C.1 Engineering Data

Data Curation Pilot 2013: Engineering Data

Dr. David Levinson is a faculty member in the civil engineering department. His primary research interest is to link the social and economic aspects of transportation networks with the physical infrastructure of transportation networks.

Type of Data

Dr. Levinson submitted the following dataset to be curated for reuse in the library. These data are the digitized and geocoded 1958 Twin Cities Land Use Map originally authored by the Twin Cities Metropolitan Planning Commission. In order to create the GIS data the print land use maps are scanned and map features are digitally "traced" to transform the image into digital GIS information. This creation took place with the significant help of his graduate student, Wei Chen and the resulting file types include the GIS files (.shp) and the scanned maps (image files and PDFs). (Note: Levinson actually contributed 5 digital map/GIS datasets, but only the 1958 map will be referenced here and total ~162MB in size).

Metadata

Levinson, on submission, choose to fill out the following metadata fields.

Title:	1958 Twin Cities Land Use Map from Twin Cities Metropolitan Planning Commission
Creator(s):	Wei Chen, David Levinson
Date:	2003-04-28
Abstract:	High-quality GIS land use maps for the Twin Cities Metropolitan Area for 1958 that were developed from paper maps (no GIS version existed previously).
Grants:	Minnesota Department of Transportation Grant: If They Come, Will You Build It? Urban Transportation Network Growth Models. MNDOT Report 2003-37. Contract #: (c) 81655 (wo) 8. Reports available at: http://nexus.umn.edu/Projects/IfTheyCome/IfTheyComeWillYouBuildIt.pdf http://www.cts.umn.edu/Publications/ResearchReports/reportdetail.html?id=686

Related Articles: Levinson, David, and Wei Chen (2007) "Area Based Models of New Highway Route Growth." ASCE Journal of Urban Planning and Development 133(4) 250-254. http://nexus.umn.edu/Papers/AreaBasedNetworkGrowth.pdf

Levinson, David and Wei Chen (2005) "Paving New Ground" in Access to Destinations (ed. David Levinson and Kevin Krizek) Elsevier Publishers. http://nexus.umn.edu/Papers/PavingNewGround.pdf

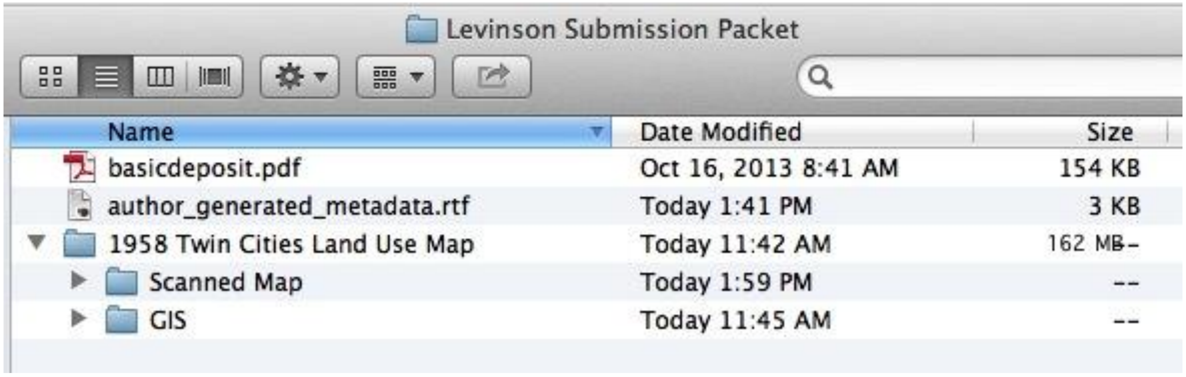
Additional Information

On submission, Levinson shared some information about the dataset with you via phone:

- These data are of public value (no sharing concerns) and have unique value as they are difficult and expensive to recreate. However they have never been widely available, just upon request (which he received occasionally via email).
- Levinson feels that they own the data and has signed a deposit agreement for the library.
- Documentation will need to be created. Some methodology can be found in Wei Chen’s PhD dissertation and resulting article. Levinson notes in a follow-up email that “Each feature in each shapefile has a LEVEL attribute with a low (<10) integer value. I think this indicates the imputed construction date?”
- There may be existing repositories for this type of data, but he is not sure. Open GeoPortal, GeoCommons? ICSPR (?) NREL
- ESRI or other GIS tool is needed to open most of the files. For preservation, .shp is becoming the standard.

Submission Information Package (SIP):

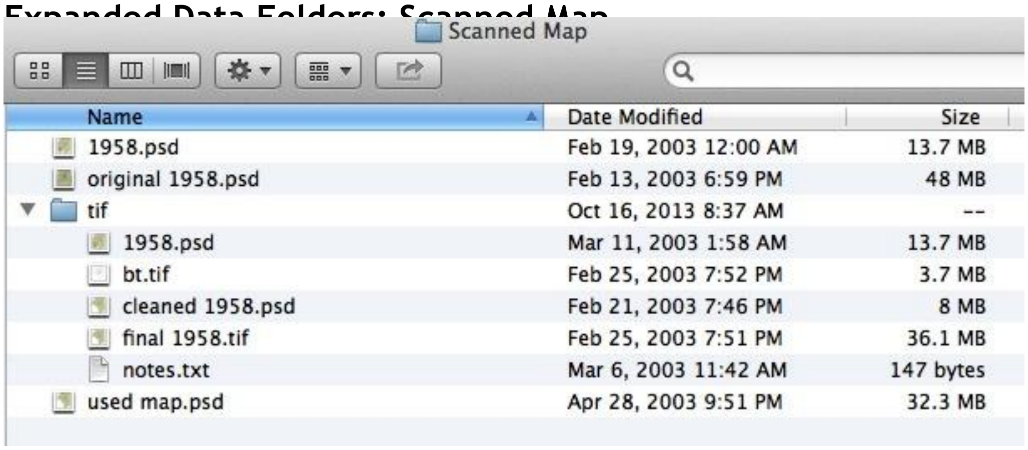
Here are all the files received from Dr. Levinson, including the metadata information that is written above, called author_generated_metadata.rtf.



Name	Date Modified	Size
basicdeposit.pdf	Oct 16, 2013 8:41 AM	154 KB
author_generated_metadata.rtf	Today 1:41 PM	3 KB
▼ 1958 Twin Cities Land Use Map	Today 11:42 AM	162 MB
▶ Scanned Map	Today 1:59 PM	--
▶ GIS	Today 11:45 AM	--

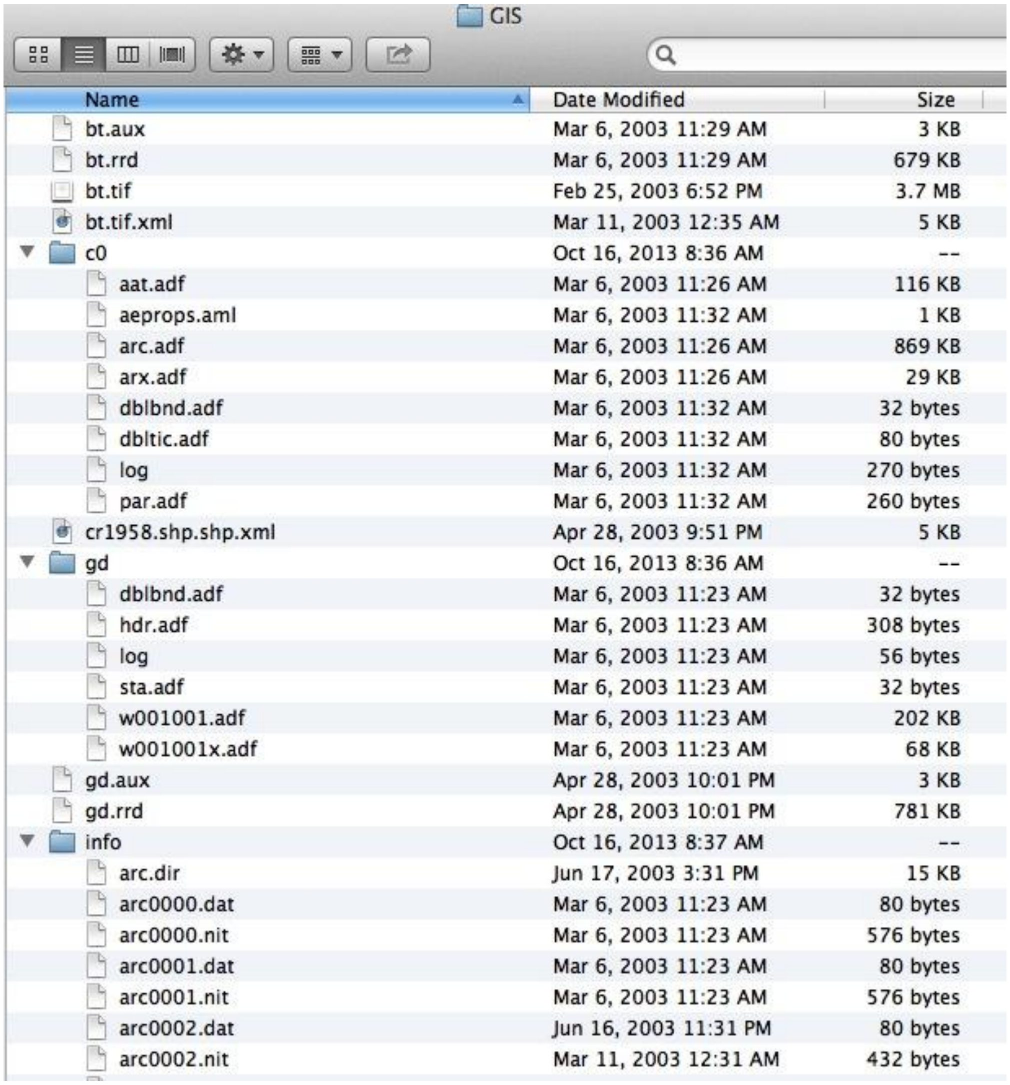
Inside each of the folders under “1958 Twin Cities Land Use Map” are the following files that total around 162MB.

Expanded Data Folders: Scanned Map



Name	Date Modified	Size
1958.psd	Feb 19, 2003 12:00 AM	13.7 MB
original 1958.psd	Feb 13, 2003 6:59 PM	48 MB
▼ tif	Oct 16, 2013 8:37 AM	--
1958.psd	Mar 11, 2003 1:58 AM	13.7 MB
bt.tif	Feb 25, 2003 7:52 PM	3.7 MB
cleaned 1958.psd	Feb 21, 2003 7:46 PM	8 MB
final 1958.tif	Feb 25, 2003 7:51 PM	36.1 MB
notes.txt	Mar 6, 2003 11:42 AM	147 bytes
used map.psd	Apr 28, 2003 9:51 PM	32.3 MB

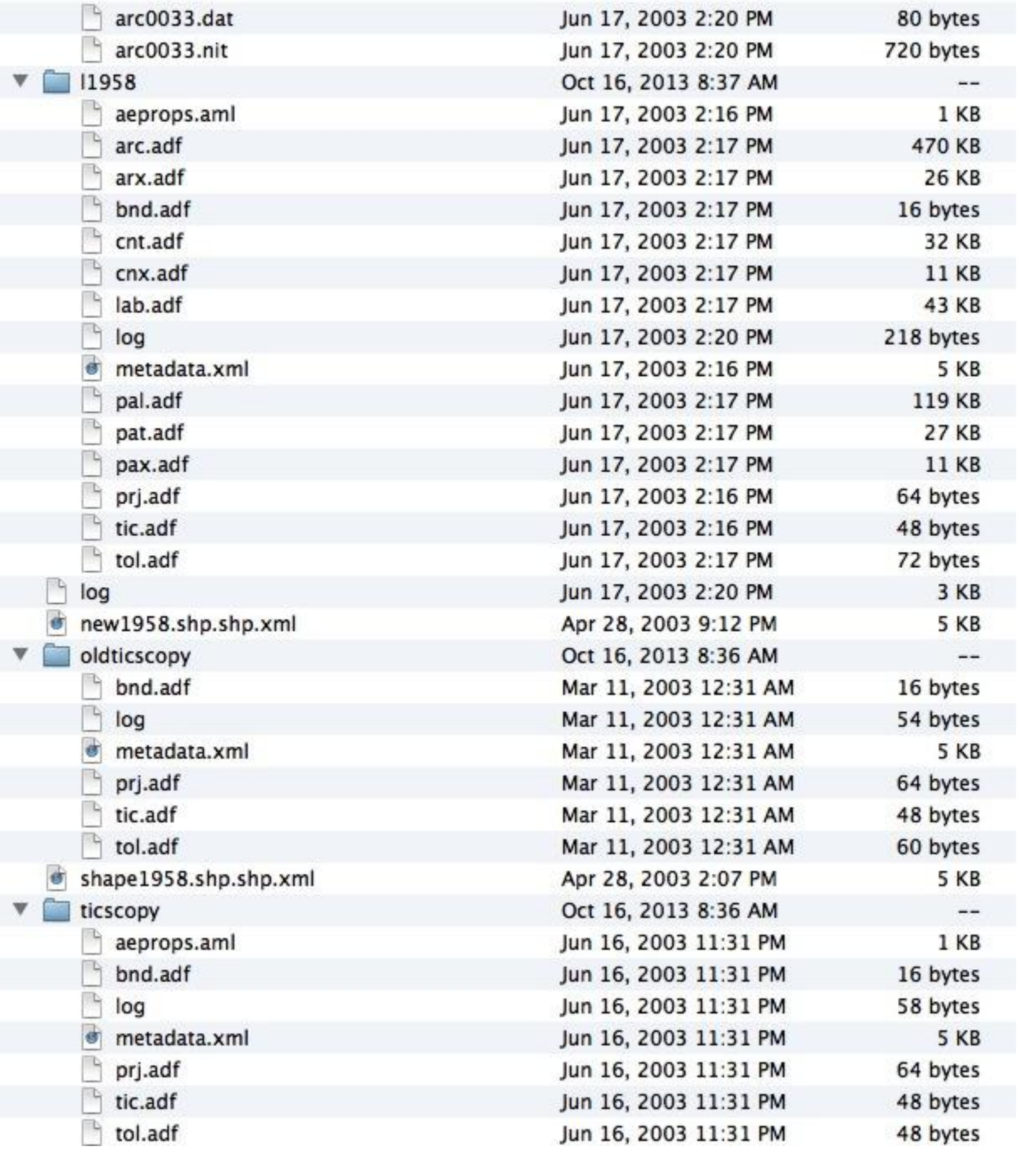
Expanded Data Folder: GIS



Name	Date Modified	Size
bt.aux	Mar 6, 2003 11:29 AM	3 KB
bt.rrd	Mar 6, 2003 11:29 AM	679 KB
bt.tif	Feb 25, 2003 6:52 PM	3.7 MB
bt.tif.xml	Mar 11, 2003 12:35 AM	5 KB
▼ c0	Oct 16, 2013 8:36 AM	--
aat.adf	Mar 6, 2003 11:26 AM	116 KB
aeprops.aml	Mar 6, 2003 11:32 AM	1 KB
arc.adf	Mar 6, 2003 11:26 AM	869 KB
arx.adf	Mar 6, 2003 11:26 AM	29 KB
dblbnd.adf	Mar 6, 2003 11:32 AM	32 bytes
dbltic.adf	Mar 6, 2003 11:32 AM	80 bytes
log	Mar 6, 2003 11:32 AM	270 bytes
par.adf	Mar 6, 2003 11:32 AM	260 bytes
cr1958.shp.shp.xml	Apr 28, 2003 9:51 PM	5 KB
▼ gd	Oct 16, 2013 8:36 AM	--
dblbnd.adf	Mar 6, 2003 11:23 AM	32 bytes
hdr.adf	Mar 6, 2003 11:23 AM	308 bytes
log	Mar 6, 2003 11:23 AM	56 bytes
sta.adf	Mar 6, 2003 11:23 AM	32 bytes
w001001.adf	Mar 6, 2003 11:23 AM	202 KB
w001001x.adf	Mar 6, 2003 11:23 AM	68 KB
gd.aux	Apr 28, 2003 10:01 PM	3 KB
gd.rrd	Apr 28, 2003 10:01 PM	781 KB
▼ info	Oct 16, 2013 8:37 AM	--
arc.dir	Jun 17, 2003 3:31 PM	15 KB
arc0000.dat	Mar 6, 2003 11:23 AM	80 bytes
arc0000.nit	Mar 6, 2003 11:23 AM	576 bytes
arc0001.dat	Mar 6, 2003 11:23 AM	80 bytes
arc0001.nit	Mar 6, 2003 11:23 AM	576 bytes
arc0002.dat	Jun 16, 2003 11:31 PM	80 bytes
arc0002.nit	Mar 11, 2003 12:31 AM	432 bytes

…. (see next page)

…. (Note files arc0003 - arc0032 are not shown for display purposes)



```
        arc0033.dat             Jun 17, 2003 2:20 PM      80 bytes
        arc0033.nit             Jun 17, 2003 2:20 PM     720 bytes
  ▼   l1958                     Oct 16, 2013 8:37 AM            --
        aeprops.aml             Jun 17, 2003 2:16 PM          1 KB
        arc.adf                 Jun 17, 2003 2:17 PM        470 KB
        arx.adf                 Jun 17, 2003 2:17 PM         26 KB
        bnd.adf                 Jun 17, 2003 2:17 PM      16 bytes
        cnt.adf                 Jun 17, 2003 2:17 PM         32 KB
        cnx.adf                 Jun 17, 2003 2:17 PM         11 KB
        lab.adf                 Jun 17, 2003 2:17 PM         43 KB
        log                     Jun 17, 2003 2:20 PM     218 bytes
        metadata.xml            Jun 17, 2003 2:16 PM          5 KB
        pal.adf                 Jun 17, 2003 2:17 PM        119 KB
        pat.adf                 Jun 17, 2003 2:17 PM         27 KB
        pax.adf                 Jun 17, 2003 2:17 PM         11 KB
        prj.adf                 Jun 17, 2003 2:16 PM      64 bytes
        tic.adf                 Jun 17, 2003 2:16 PM      48 bytes
        tol.adf                 Jun 17, 2003 2:17 PM      72 bytes
      log                       Jun 17, 2003 2:20 PM          3 KB
      new1958.shp.shp.xml       Apr 28, 2003 9:12 PM          5 KB
  ▼   oldticscopy               Oct 16, 2013 8:36 AM            --
        bnd.adf                 Mar 11, 2003 12:31 AM     16 bytes
        log                     Mar 11, 2003 12:31 AM     54 bytes
        metadata.xml            Mar 11, 2003 12:31 AM         5 KB
        prj.adf                 Mar 11, 2003 12:31 AM     64 bytes
        tic.adf                 Mar 11, 2003 12:31 AM     48 bytes
        tol.adf                 Mar 11, 2003 12:31 AM     60 bytes
      shape1958.shp.shp.xml     Apr 28, 2003 2:07 PM          5 KB
  ▼   ticscopy                  Oct 16, 2013 8:36 AM            --
        aeprops.aml             Jun 16, 2003 11:31 PM         1 KB
        bnd.adf                 Jun 16, 2003 11:31 PM     16 bytes
        log                     Jun 16, 2003 11:31 PM     58 bytes
        metadata.xml            Jun 16, 2003 11:31 PM         5 KB
        prj.adf                 Jun 16, 2003 11:31 PM     64 bytes
        tic.adf                 Jun 16, 2003 11:31 PM     48 bytes
        tol.adf                 Jun 16, 2003 11:31 PM     48 bytes
```

C.2 Health Sciences Data

Data Curation Pilot 2013: Health Sciences Data

Dr. James Hodges is a biostatistician in the School of Public Health. His primary research interest is oral-health research, infectious diseases, neurology, and kidney disease.

Type of Data

Dr. Hodges submitted the following dataset to be curated for reuse in the library. These data are histographic periodontal data, or large-scale studies that measure participant's teeth over time. The primary author of the data is Bryan Michalowicz from the UMN School of Dentistry, who was PI of the grant that Hodges collaborated on. A public-use version of the dataset was created in November 2009, but has never been released in a repository. The file format for the raw data is in .csv and .txt format since the large 5MB file is not operable in MS Excel. The personal-level (proceeded) dataset is in MS Excel and 1.7MBs. Both are deidentified for public use and total ~ 12.8MB in size.

Metadata

Hodges, on submission, choose to fill out the following metadata fields.

Title: Public-Use Data from the Obstetrics and Periodontal Therapy (OPT) Study, a randomized trial of periodontal therapy to prevent pre-term birth

Creator(s): James S. Hodges -- author of record in Digital Conservancy Agreement
Bryan S. Michalowicz -- Principal Investigator of the OPT Study

Date: 2009-11-02

Coverage,
Temporal: 2003-2006

Abstract: The OPT Study was a multi-center randomized, single-blind (examiners) controlled clinical trial testing whether mechanical periodontal therapy (scaling and planning) in pregnant women at risk for premature birth reduced the extent or severity of premature birth. OPT found that periodontal therapy does not reduce the number or timing of premature births. Nonetheless, this public-use dataset is of interest in that it provides natural and clinical histories of the periodontal status of pregnant women with treated and untreated periodontal disease.

Data include birth outcomes (including gestational age, birth weight, presence of congenital anomalies, and 1 and 5 minute APGAR scores), baseline characteristics (including previous pregnancy outcomes), periodontal therapy and essential dental care delivered as part of the study, maternal conditions during pregnancy, and the following items for three visits between the end of the first trimester and delivery: clinical periodontal measurements (pocket depth, attachment loss, and bleeding on probing at 6 sites per tooth; site-specific data and several common person-level summaries), medications, dental plaque levels of 8 bacterial species, levels of serum antibodies for the same 8 bacterial species, and serum levels of 8 inflammatory markers or mediators. The OPT Study's Manual of Procedures (Version 1) is available as part of this package. Version changes during the course of the study were rare and affected very few data items (mostly the data describing study periodontal therapy). The public-use dataset includes the version of these data items used in the main and secondary papers.

The OPT Study team published the main paper in 2006 in the New England Journal of Medicine and has published 8 secondary papers. The dataset was made available in 2009 but has not previously been available online because the sponsoring agency had no facility for making it available.

Grants: The data in these files were collected during the OPT Study funded by the National Institute of Dental and Craniofacial Research, grant number DE014338.

Keywords: Periodontal therapy, Pre-term birth, Randomized controlled trial, Periodontal measurements, Pregnant women

Related Articles: (main paper) Michalowicz BS, Hodges JS, DiAngelis AJ, Lupo VR, Novak MJ, Ferguson JE, Buchanan W, Bofill J, Papapanou PN, Mitchell DA, Matseoane S, Tschida PA; OPT Study. Treatment of periodontal disease and the risk of preterm birth. N Engl J Med. 2006 Nov 2;355(18):1885-94.

(Note: 8 secondary papers not listed here for presentation purposes)

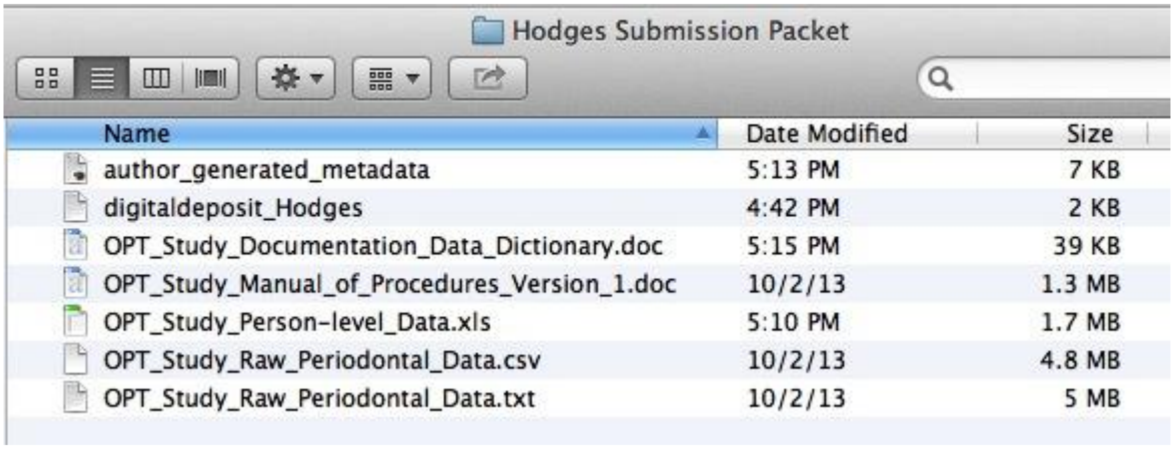
Additional Information

On submission, Hodges shared some information about the dataset with you via phone:

- These data are a public-use dataset but the funding institute, the National Institute of Dental & Craniofacial Research, has no facility for hosting such datasets even though their standard contract requires that such a public-use dataset be preserved.
- To host the data, Dr. Hodges explains that “it sits on my computer and if anybody asks for it, I send them the dataset and the modest documentation. This has been 2 people so far. It'd be more available if it were in a repository of similar datasets.”
- Hodges feels that he owns the data and has signed a deposit agreement for the library.
- Documentation was added due to participating in this study. Those files include a Data Dictionary and a Manual of Procedures.

Submission Information Package (SIP):

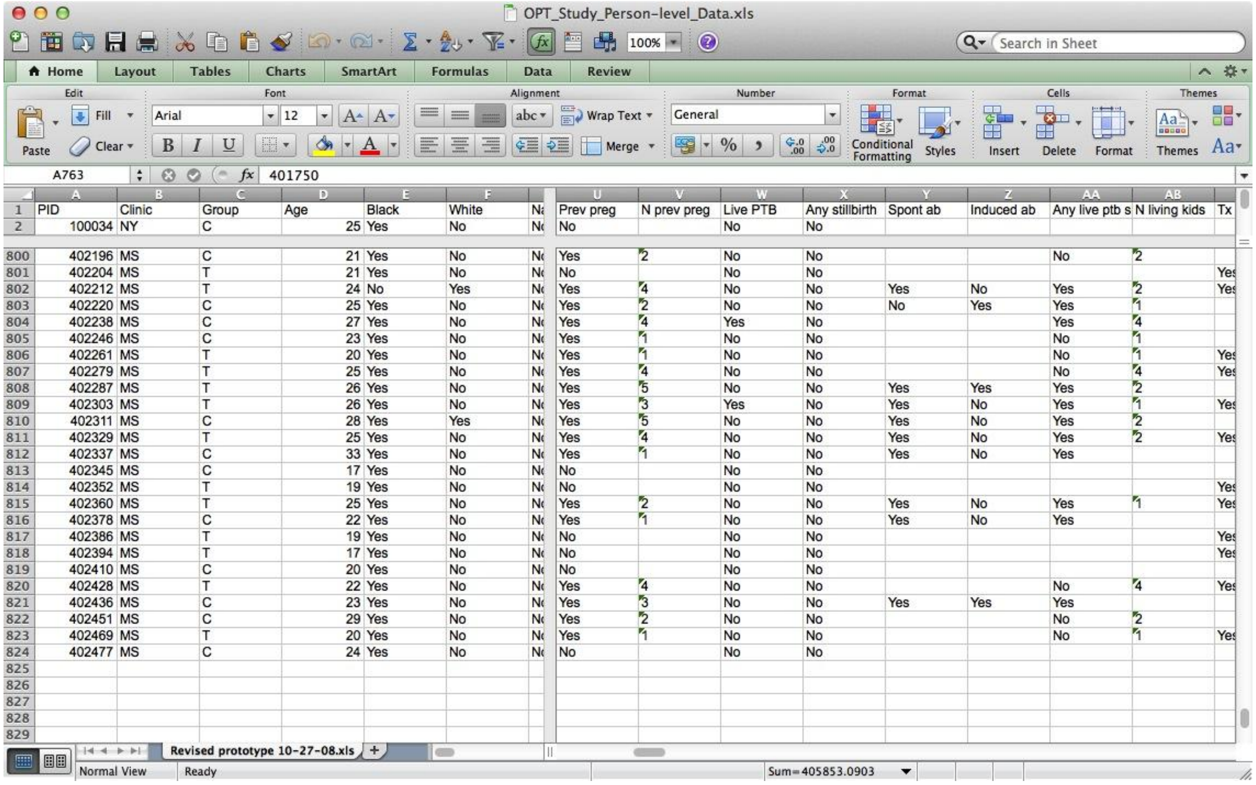
Here are all the files received from Dr. Hodges, including the metadata information that is written above, called author_generated_metadata.rtf.



Hodges Submission Packet

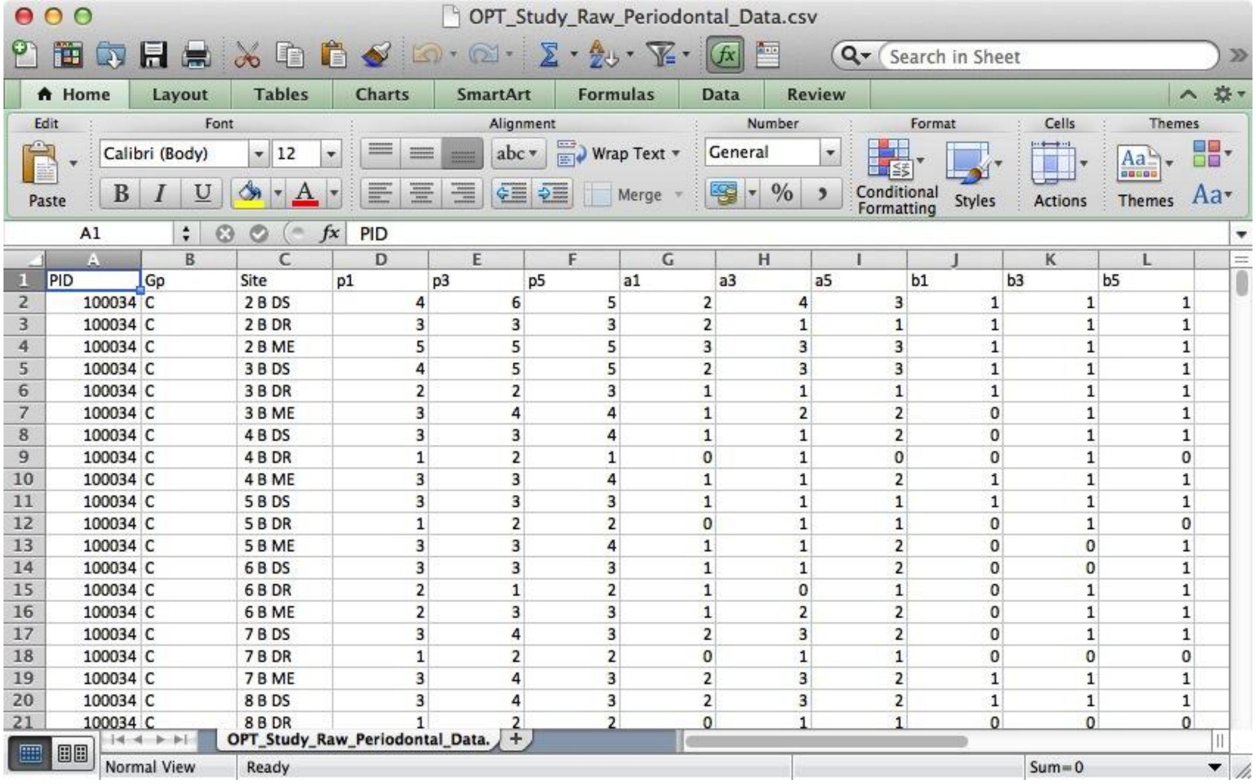
Name	Date Modified	Size
author_generated_metadata	5:13 PM	7 KB
digitaldeposit_Hodges	4:42 PM	2 KB
OPT_Study_Documentation_Data_Dictionary.doc	5:15 PM	39 KB
OPT_Study_Manual_of_Procedures_Version_1.doc	10/2/13	1.3 MB
OPT_Study_Person-level_Data.xls	5:10 PM	1.7 MB
OPT_Study_Raw_Periodontal_Data.csv	10/2/13	4.8 MB
OPT_Study_Raw_Periodontal_Data.txt	10/2/13	5 MB

Screenshot of OPT_Study_Person-level_Data.xls

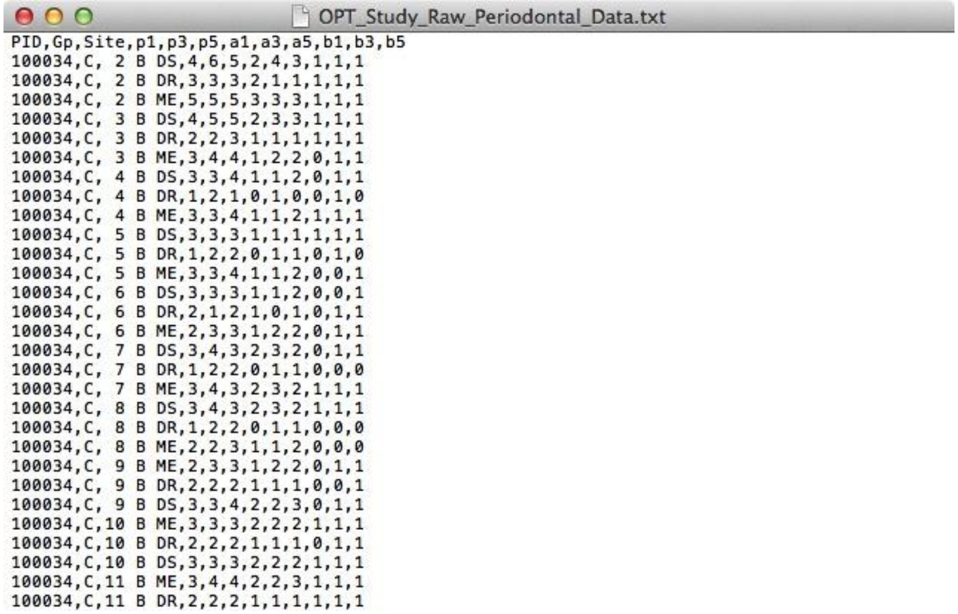


	A	B	C	D	E	F		U	V	W	X	Y	Z	AA	AB	
1	PID	Clinic	Group	Age	Black	White	Na	Prev preg	N prev preg	Live PTB	Any stillbirth	Spont ab	Induced ab	Any live ptb s	N living kids	Tx
2	100034	NY	C	25	Yes	No	No	No		No	No					
800	402196	MS	C	21	Yes	No	No	Yes	2	No	No			No	2	
801	402204	MS	T	21	Yes	No	No	No		No	No					Yes
802	402212	MS	T	24	No	Yes	No	Yes	4	No	No	Yes	No	Yes	2	Yes
803	402220	MS	C	25	Yes	No	No	Yes	2	No	No	No	Yes	Yes	1	
804	402238	MS	C	27	Yes	No	No	Yes	4	Yes	No			Yes	4	
805	402246	MS	C	23	Yes	No	No	Yes	1	No	No			No	1	
806	402261	MS	T	20	Yes	No	No	Yes	1	No	No			No	1	Yes
807	402279	MS	T	25	Yes	No	No	Yes	4	No	No			No	4	Yes
808	402287	MS	T	26	Yes	No	No	Yes	5	No	No	Yes	Yes	Yes	2	
809	402303	MS	T	26	Yes	No	No	Yes	3	Yes	No	Yes	No	Yes	1	Yes
810	402311	MS	C	28	Yes	Yes	No	Yes	5	No	No	Yes	No	Yes	2	
811	402329	MS	T	25	Yes	No	No	Yes	4	No	No	Yes	No	Yes	2	Yes
812	402337	MS	C	33	Yes	No	No	Yes	1	No	No	Yes	No	Yes		
813	402345	MS	C	17	Yes	No	No	No		No	No					
814	402352	MS	T	19	Yes	No	No	No		No	No					Yes
815	402360	MS	T	25	Yes	No	No	Yes	2	No	No	Yes	No	Yes	1	Yes
816	402378	MS	C	22	Yes	No	No	Yes	1	No	No	Yes	No	Yes		
817	402386	MS	T	19	Yes	No	No	No		No	No					Yes
818	402394	MS	T	17	Yes	No	No	No		No	No					Yes
819	402410	MS	C	20	Yes	No	No	No		No	No					
820	402428	MS	T	22	Yes	No	No	Yes	4	No	No			No	4	Yes
821	402436	MS	C	23	Yes	No	No	Yes	3	No	No	Yes	Yes	Yes		
822	402451	MS	C	29	Yes	No	No	Yes	2	No	No			No	2	
823	402469	MS	T	20	Yes	No	No	Yes	1	No	No			No	1	Yes
824	402477	MS	C	24	Yes	No	No	No		No	No					
825																
826																
827																
828																
829																

Screenshot of OPT Study Raw Periodontal Data (.csv and .txt)

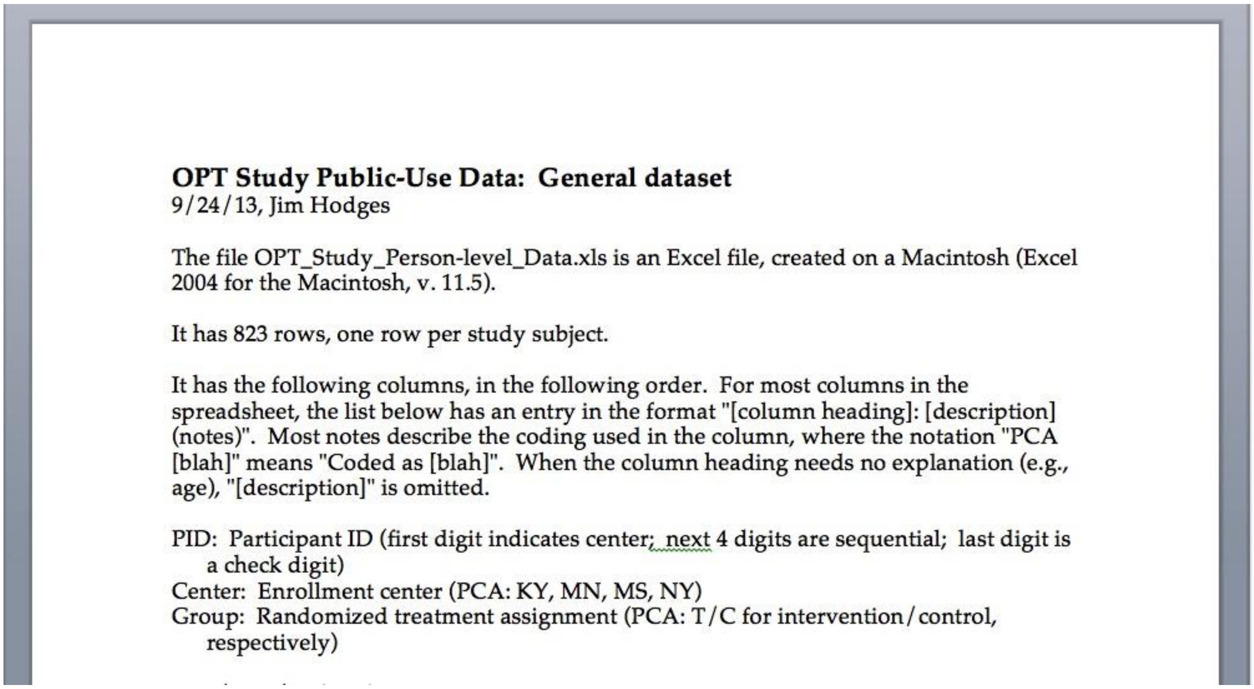


	A	B	C	D	E	F	G	H	I	J	K	L
1	PID	Gp	Site	p1	p3	p5	a1	a3	a5	b1	b3	b5
2	100034	C	2 B DS	4	6	5	2	4	3	1	1	1
3	100034	C	2 B DR	3	3	3	2	1	1	1	1	1
4	100034	C	2 B ME	5	5	5	3	3	3	1	1	1
5	100034	C	3 B DS	4	5	5	2	3	3	1	1	1
6	100034	C	3 B DR	2	2	3	1	1	1	1	1	1
7	100034	C	3 B ME	3	4	4	1	2	2	0	1	1
8	100034	C	4 B DS	3	3	4	1	1	2	0	1	1
9	100034	C	4 B DR	1	2	1	0	1	0	0	1	0
10	100034	C	4 B ME	3	3	4	1	1	2	1	1	1
11	100034	C	5 B DS	3	3	3	1	1	1	1	1	1
12	100034	C	5 B DR	1	2	2	0	1	1	0	1	0
13	100034	C	5 B ME	3	3	4	1	1	2	0	0	1
14	100034	C	6 B DS	3	3	3	1	1	2	0	0	1
15	100034	C	6 B DR	2	1	2	1	0	1	0	1	1
16	100034	C	6 B ME	2	3	3	1	2	2	0	1	1
17	100034	C	7 B DS	3	4	3	2	3	2	0	1	1
18	100034	C	7 B DR	1	2	2	0	1	1	0	0	0
19	100034	C	7 B ME	3	4	3	2	3	2	1	1	1
20	100034	C	8 B DS	3	4	3	2	3	2	1	1	1
21	100034	C	8 B DR	1	2	2	0	1	1	0	0	0



```
PID,Gp,Site,p1,p3,p5,a1,a3,a5,b1,b3,b5
100034,C, 2 B DS,4,6,5,2,4,3,1,1,1
100034,C, 2 B DR,3,3,3,2,1,1,1,1,1
100034,C, 2 B ME,5,5,5,3,3,3,1,1,1
100034,C, 3 B DS,4,5,5,2,3,3,1,1,1
100034,C, 3 B DR,2,2,3,1,1,1,1,1,1
100034,C, 3 B ME,3,4,4,1,2,2,0,1,1
100034,C, 4 B DS,3,3,4,1,1,2,0,1,1
100034,C, 4 B DR,1,2,1,0,1,0,0,1,0
100034,C, 4 B ME,3,3,4,1,1,2,1,1,1
100034,C, 5 B DS,3,3,3,1,1,1,1,1,1
100034,C, 5 B DR,1,2,2,0,1,1,0,1,0
100034,C, 5 B ME,3,3,4,1,1,2,0,0,1
100034,C, 6 B DS,3,3,3,1,1,2,0,0,1
100034,C, 6 B DR,2,1,2,1,0,1,0,1,1
100034,C, 6 B ME,2,3,3,1,2,2,0,1,1
100034,C, 7 B DS,3,4,3,2,3,2,0,1,1
100034,C, 7 B DR,1,2,2,0,1,1,0,0,0
100034,C, 7 B ME,3,4,3,2,3,2,1,1,1
100034,C, 8 B DS,3,4,3,2,3,2,1,1,1
100034,C, 8 B DR,1,2,2,0,1,1,0,0,0
100034,C, 8 B ME,2,2,3,1,1,2,0,0,0
100034,C, 9 B ME,2,3,3,1,2,2,0,1,1
100034,C, 9 B DR,2,2,2,1,1,1,0,0,1
100034,C, 9 B DS,3,3,4,2,2,3,0,1,1
100034,C,10 B ME,3,3,3,2,2,2,1,1,1
100034,C,10 B DR,2,2,2,1,1,1,0,1,1
100034,C,10 B DS,3,3,3,2,2,2,1,1,1
100034,C,11 B ME,3,4,4,2,2,3,1,1,1
100034,C,11 B DR,2,2,2,1,1,1,1,1,1
```

Screenshot of OPT_Study_Documentation_Data_Dictionary.doc



OPT Study Public-Use Data: General dataset
9/24/13, Jim Hodges

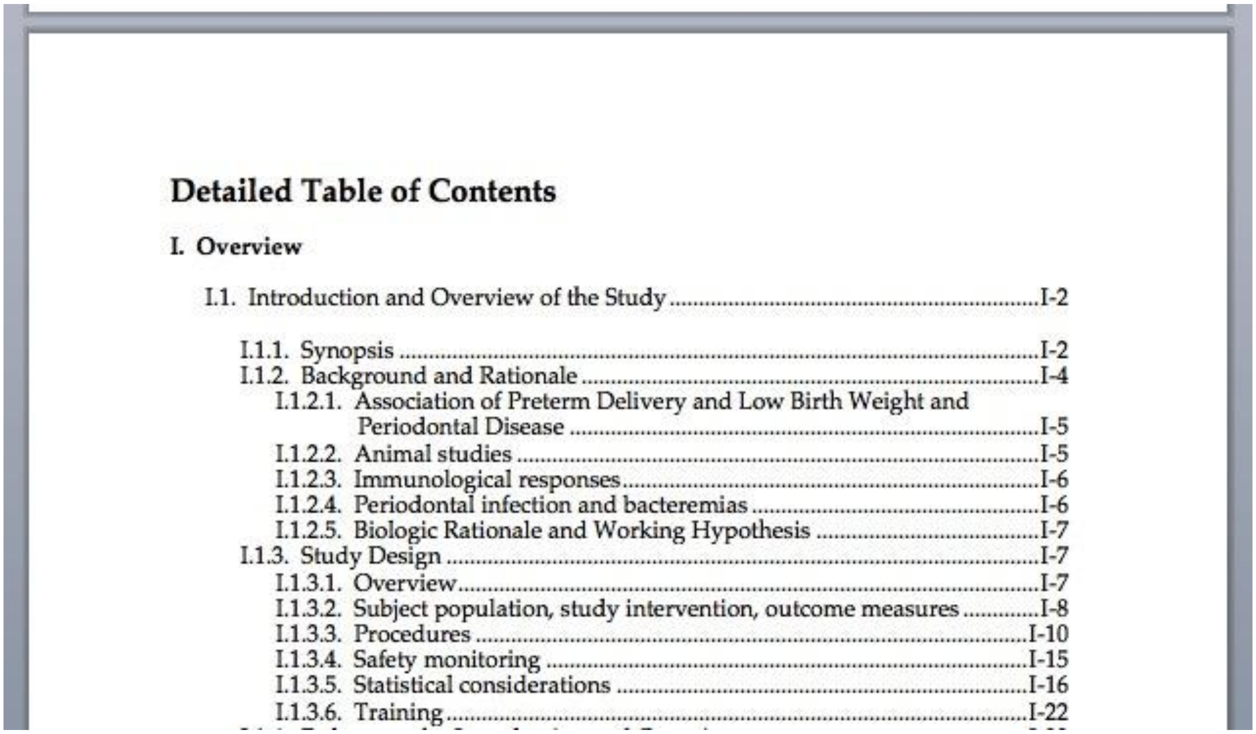
The file OPT_Study_Person-level_Data.xls is an Excel file, created on a Macintosh (Excel 2004 for the Macintosh, v. 11.5).

It has 823 rows, one row per study subject.

It has the following columns, in the following order. For most columns in the spreadsheet, the list below has an entry in the format "[column heading]: [description] (notes)". Most notes describe the coding used in the column, where the notation "PCA [blah]" means "Coded as [blah]". When the column heading needs no explanation (e.g., age), "[description]" is omitted.

PID: Participant ID (first digit indicates center; next 4 digits are sequential; last digit is a check digit)
Center: Enrollment center (PCA: KY, MN, MS, NY)
Group: Randomized treatment assignment (PCA: T/C for intervention/control, respectively)

Screenshot of OPT_Study_Manual_of_Procedures_Version_1.doc



Detailed Table of Contents

I. Overview

C.3 Interdisciplinary Data

Data Curation Pilot 2013: Interdisciplinary Data

Dr. Bill Arnold is a associate department head in the civil engineering department. His discipline is "Water Chemistry," or the chemistry of natural and engineered aquatic systems, and his research focus is to predict contaminant rates in natural aquatic systems and to design remediation technologies to treat contaminated waters.

Type of Data

Dr. Arnold submitted the following dataset to be curated for reuse in the library. These data are histographic samples of chemical concentrations in sediment cores. The data was completed with the significant help of Arnold's graduate student, Cale T. Anger, who wrote a dissertation using the data (Note: Arnold included Anger's thesis draft along with the data as documentation). The excel files are small (100-300kb) yet complex with multiple sheets per worksheet, many charts and graphs, and use colored cells to distinguish aspects of the data. These data have not been previously released but must be to meet NSF data sharing requirements as they may be of great interest to stakeholders such as regulators, environmental consultations, and other researchers.

Metadata

Arnold, on submission, choose to fill out the following metadata fields.

Title: Triclosan, chlorinated triclosan derivative, and dioxin levels in Minnesota lakes

Creator(s): Cale T. Anger, Charles Sueper, Dylan J. Blumentritt, Kristopher Mcneill, Daniel R. Engstrom, William A. Arnold

Keywords: Sediment, contaminants, triclosan, pharmaceuticals, pollution, dioxins

Related Article(s): Cale T. Anger, Charles Sueper, Dylan J. Blumentritt, Kristopher McNeill, Daniel R. Engstrom, and William A. Arnold. (2013). Quantification of Triclosan, Chlorinated Triclosan Derivatives, and their Dioxin Photoproducts in Lacustrine Sediment Cores. *Environmental Science & Technology* 2013 47 (4), 1833-1843. dx.doi.org/10.1021/es3045289

Date: 2012-09-01

Coverage, Temporal: Data collected from 2010-2012

Funder/
Grant: Funding for this project was provided by the Minnesota Environment and Natural Resources Trust Fund as recommended by the Legislative-Citizen Commission on Minnesota Resources and the National Science Foundation (CBET 0967163)

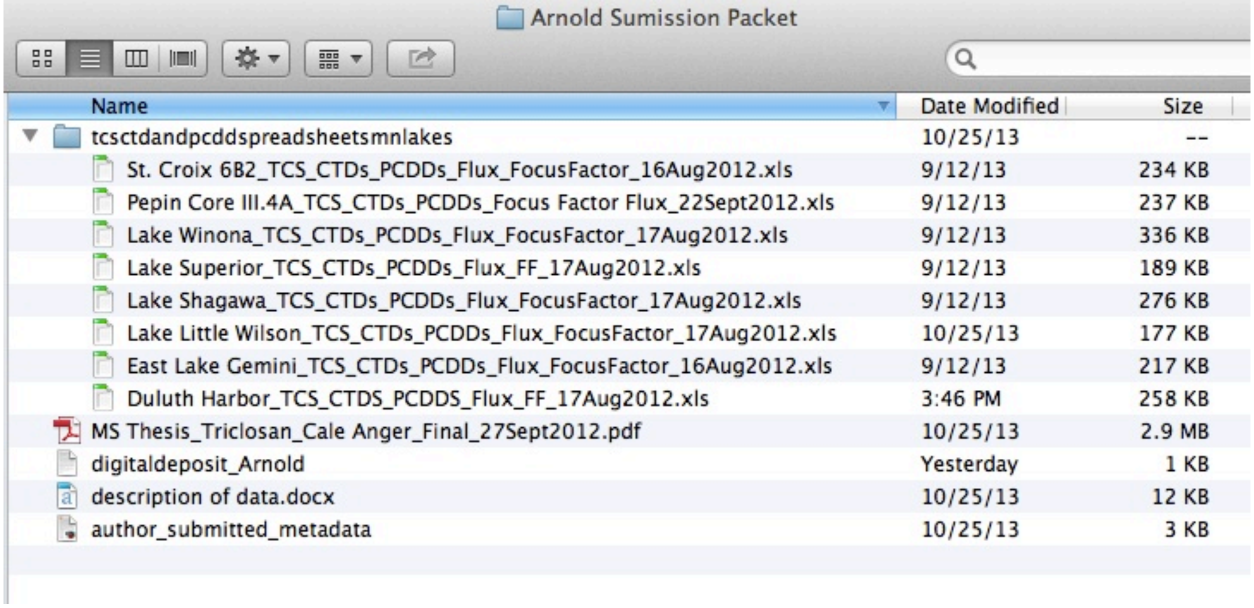
Additional Information

On submission, Arnold shared some information about the dataset with you via phone:

- The raw data used to produce the excel files are generated by instruments that are accessed only with proprietary software (and thus the files are not interoperable, even with similar instruments). Nevertheless, for the long-term the raw instrument data should be archived as well. Repositories like Earth Cube might be a good place for this type of data, but Arnold is not sure about the software needed to open the files and if there is a non-proprietary substitute.
- Documentation for the data was included in the graduate student's thesis and Arnold thought is might be repurposed for documentation.
- Arnold feels that he owns the data and has signed a deposit agreement for the library.
- This data is difficult to reproduce and cost nearly 250,000 to create.

Submission Information Package (SIP):

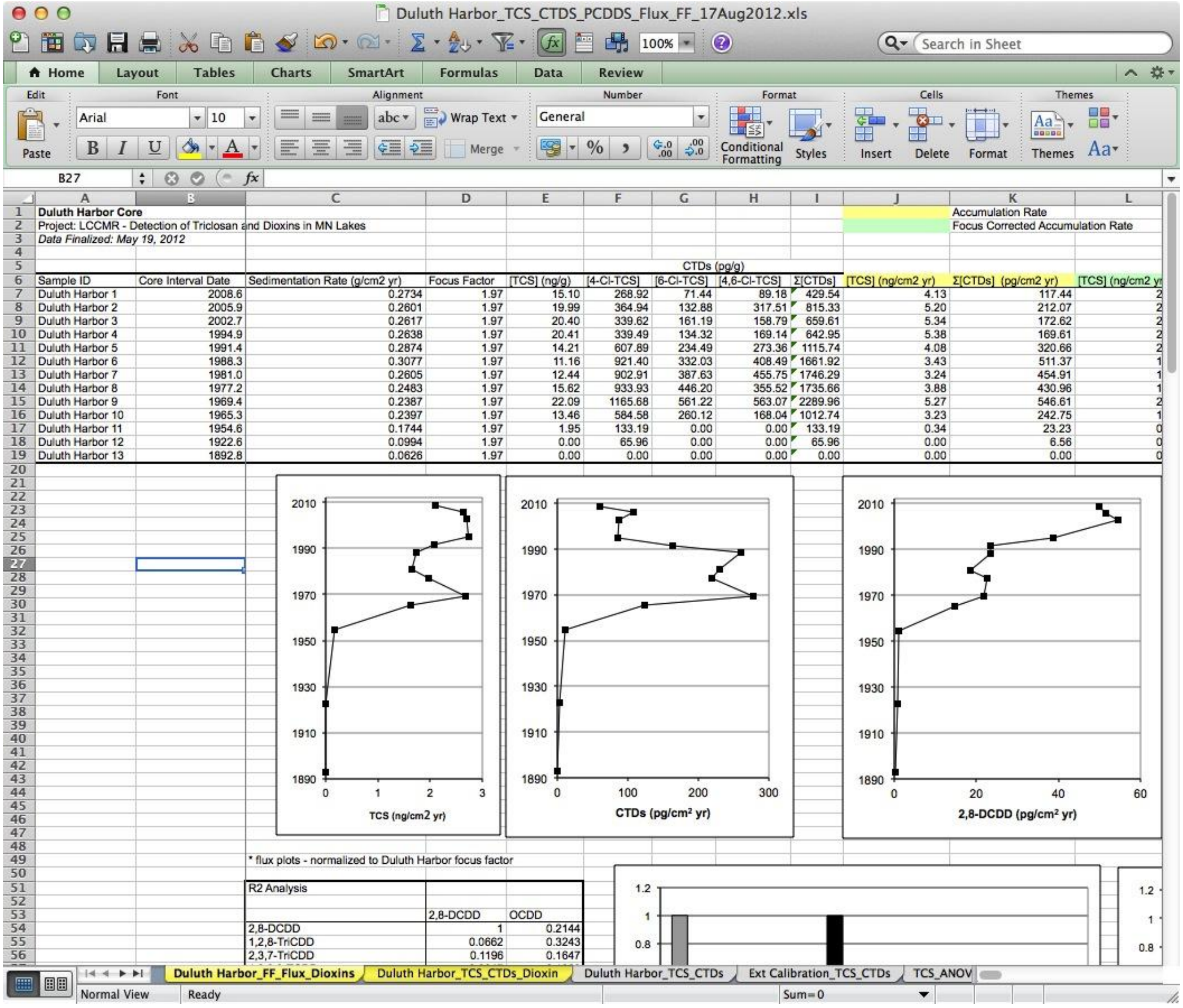
Here are all the files received from Dr. Arnold, including the metadata information that is written above, called author_generated_metadata.rtf.



Name	Date Modified	Size
▼ tcsctdandpcddspreadsheetsmnlakes	10/25/13	--
St. Croix 6B2_TCS_CTDs_PCDDs_Flux_FocusFactor_16Aug2012.xls	9/12/13	234 KB
Pepin Core III.4A_TCS_CTDs_PCDDs_Focus Factor Flux_22Sept2012.xls	9/12/13	237 KB
Lake Winona_TCS_CTDs_PCDDs_Flux_FocusFactor_17Aug2012.xls	9/12/13	336 KB
Lake Superior_TCS_CTDs_PCDDs_Flux_FF_17Aug2012.xls	9/12/13	189 KB
Lake Shagawa_TCS_CTDs_PCDDs_Flux_FocusFactor_17Aug2012.xls	9/12/13	276 KB
Lake Little Wilson_TCS_CTDs_PCDDs_Flux_FocusFactor_17Aug2012.xls	10/25/13	177 KB
East Lake Gemini_TCS_CTDs_PCDDs_Flux_FocusFactor_16Aug2012.xls	9/12/13	217 KB
Duluth Harbor_TCS_CTDS_PCDDS_Flux_FF_17Aug2012.xls	3:46 PM	258 KB
MS Thesis_Triclosan_Cale Anger_Final_27Sept2012.pdf	10/25/13	2.9 MB
digitaldeposit_Arnold	Yesterday	1 KB
description of data.docx	10/25/13	12 KB
author_submitted_metadata	10/25/13	3 KB

Screenshot of DuluthHarbor_TCS_CTDS_PCDDS_Flux_FocusFactor_16Aug2012.xls

(Note: There are 8 Lake .xls files, one is shown here as an example)



Duluth Harbor Core

Project: LCCMR - Detection of Triclosan and Dioxins in MN Lakes

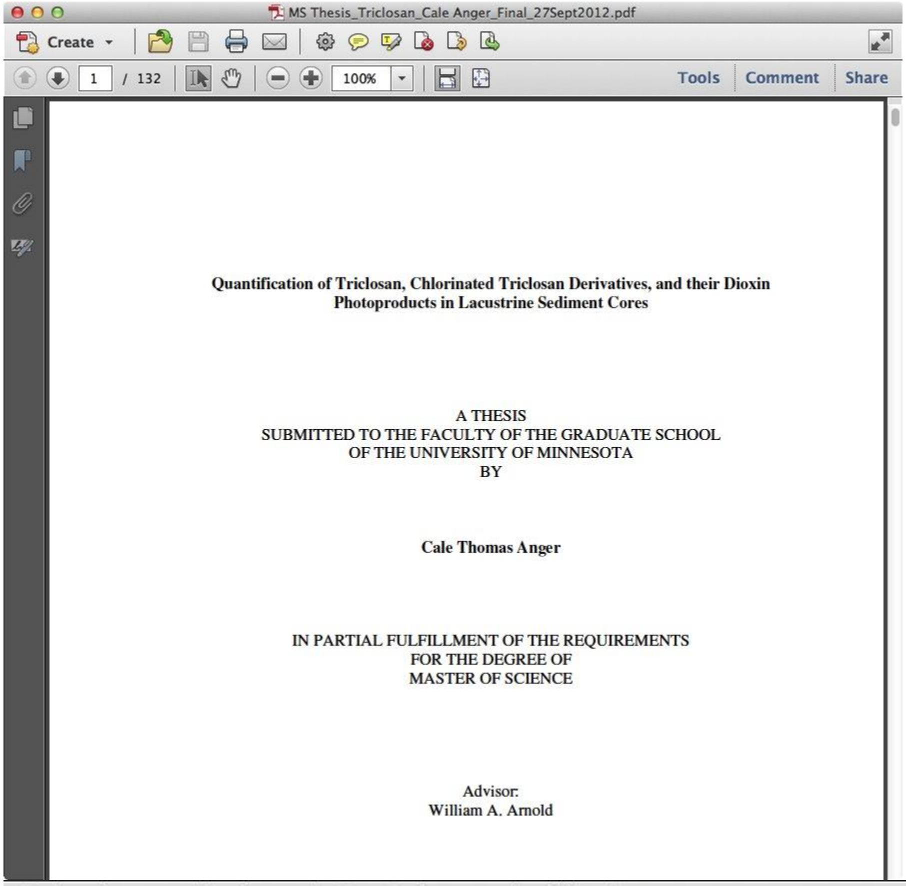
Data Finalized: May 19, 2012

										Accumulation Rate	
										Focus Corrected Accumulation Rate	
					CTDs (pg/g)						
Sample ID	Core Interval Date	Sedimentation Rate (g/cm2 yr)	Focus Factor	[TCS] (ng/g)	[4-Cl-TCS]	[6-Cl-TCS]	[4,6-Cl-TCS]	Σ[CTDs]	[TCS] (ng/cm2 yr)	Σ[CTDs] (pg/cm2 yr)	[TCS] (ng/cm2 y
Duluth Harbor 1	2008.6	0.2734	1.97	15.10	268.92	71.44	89.18	429.54	4.13	117.44	2
Duluth Harbor 2	2005.9	0.2601	1.97	19.99	364.94	132.88	317.51	815.33	5.20	212.07	2
Duluth Harbor 3	2002.7	0.2617	1.97	20.40	339.62	161.19	158.79	659.61	5.34	172.62	2
Duluth Harbor 4	1994.9	0.2638	1.97	20.41	339.49	134.32	169.14	642.95	5.38	169.61	2
Duluth Harbor 5	1991.4	0.2874	1.97	14.21	607.89	234.49	273.36	1115.74	4.08	320.66	2
Duluth Harbor 6	1988.3	0.3077	1.97	11.16	921.40	332.03	408.49	1661.92	3.43	511.37	1
Duluth Harbor 7	1981.0	0.2605	1.97	12.44	902.91	387.63	455.75	1746.29	3.24	454.91	1
Duluth Harbor 8	1977.2	0.2483	1.97	15.62	933.93	446.20	355.52	1735.66	3.88	430.96	1
Duluth Harbor 9	1969.4	0.2387	1.97	22.09	1165.68	561.22	563.07	2289.96	5.27	546.61	2
Duluth Harbor 10	1965.3	0.2397	1.97	13.46	584.58	260.12	168.04	1012.74	3.23	242.75	1
Duluth Harbor 11	1954.6	0.1744	1.97	1.95	133.19	0.00	0.00	133.19	0.34	23.23	0
Duluth Harbor 12	1922.6	0.0994	1.97	0.00	65.96	0.00	0.00	65.96	0.00	6.56	0
Duluth Harbor 13	1892.8	0.0626	1.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0

* flux plots - normalized to Duluth Harbor focus factor

R2 Analysis		
	2,8-DCDD	OCDD
2,8-DCDD	1	0.2144
1,2,8-TriCDD	0.0662	0.3243
2,3,7-TriCDD	0.1196	0.1647

Screenshot of MS Thesis_Triclosan_Cale Anger_Final_27Sept2012.pdf



Quantification of Triclosan, Chlorinated Triclosan Derivatives, and their Dioxin Photoproducts in Lacustrine Sediment Cores

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Cale Thomas Anger

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisor:
William A. Arnold

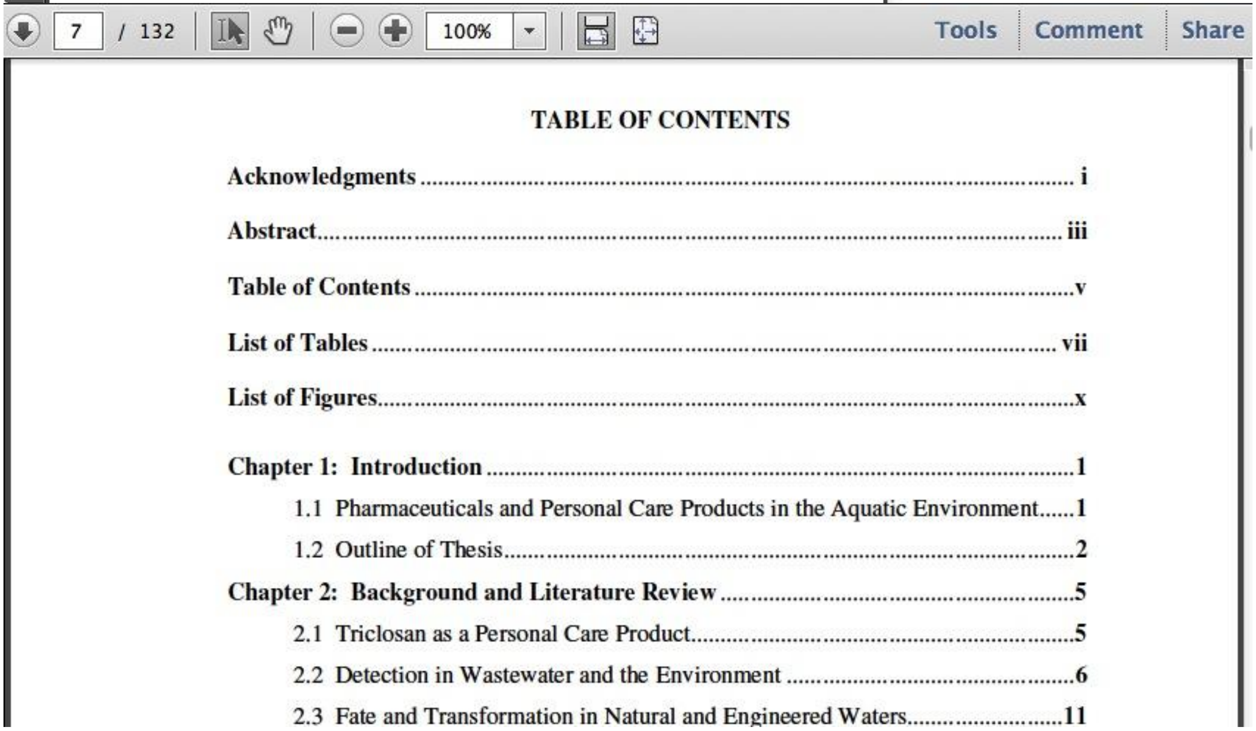
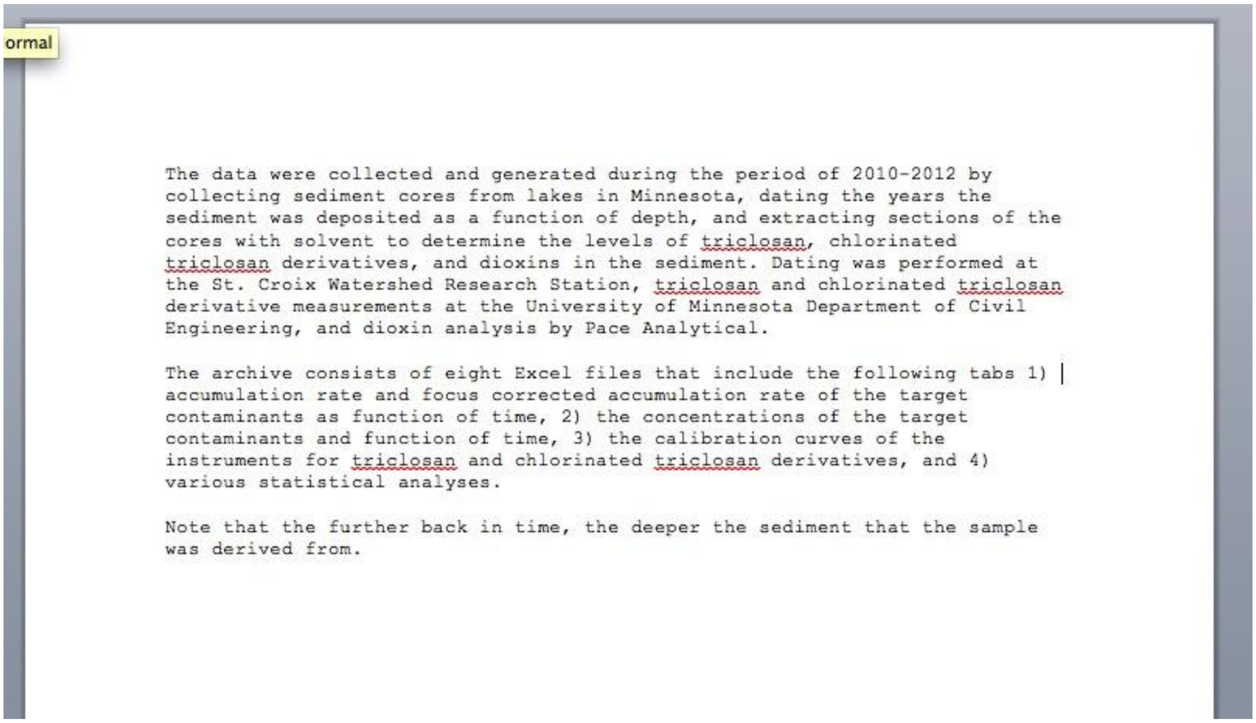


TABLE OF CONTENTS

Screenshot of description of data.doc



The data were collected and generated during the period of 2010-2012 by collecting sediment cores from lakes in Minnesota, dating the years the sediment was deposited as a function of depth, and extracting sections of the cores with solvent to determine the levels of triclosan, chlorinated triclosan derivatives, and dioxins in the sediment. Dating was performed at the St. Croix Watershed Research Station, triclosan and chlorinated triclosan derivative measurements at the University of Minnesota Department of Civil Engineering, and dioxin analysis by Pace Analytical.

The archive consists of eight Excel files that include the following tabs 1) accumulation rate and focus corrected accumulation rate of the target contaminants as function of time, 2) the concentrations of the target contaminants and function of time, 3) the calibration curves of the instruments for triclosan and chlorinated triclosan derivatives, and 4) various statistical analyses.

Note that the further back in time, the deeper the sediment that the sample was derived from.

C.4 Natural Resources Data

Data Curation Pilot 2013: Natural Resources Data

Dr. Xinyi (Lisa) Qian is a research associate in the Department of Forest Resources. She works closely with the Tourism Center, part of the UMN's Extension program, to better understand the human dimensions of natural resources and environmental management.

Type of Data

Dr. Qian submitted the following dataset to be curated for reuse in the library. These data are results from an online survey, conducted by the Tourism Center, of businesses within the industry in 2007, 2010, and 2013. With this longitudinal dimension, the data have become even more relevant and valuable to researchers. The three survey results have been combined into one SPSS (statistical analysis program) file for analysis, a .sav file, and is 37.6MB in size. The instruments were administered via the web and Qian only has access to the PDF screenshots for each year. The data has not been previously released, however reports for each year have been posted on the Tourism Centers' website.

Metadata

Qian, on submission, choose to fill out the following metadata fields.

Title: State of Sustainability Practices among Minnesota Tourism Businesses, 2007-2013

Creator(s): Xinyi (Lisa) Qian, Ingrid E. Schneider

Keywords: Sustainable tourism, Sustainability practice, Tourism businesses, Benefit, Difficulty, Energy efficiency, Waste minimization, Environmental purchasing, Air quality, Water conservation, Landscaping, Wildlife

Abstract: The dataset was used in three major ways. First, using data collected in 2013, we documented the current attitude towards sustainability practices among tourism businesses in Minnesota, particularly how they perceive the benefits and difficulties of implementing these practices. We also documented the extent of implementation of six types of sustainability practices, including energy efficiency, waste minimization, environmental purchasing, air quality, water conservation, and landscaping/wildlife. Second, we assessed whether attitude towards sustainability practices and the extent of implementing various practices changed over time (i.e., across the three surveys). Lastly, we benchmarked current level of knowledge of invasive species among Minnesota tourism businesses using data from the 2013 survey. This is the first time that the survey includes questions that assess knowledge of invasive species, providing a benchmarking opportunity.

We want to release this dataset, because there is little research that documents the extent to which different types of sustainability practices are implemented among tourism businesses in the state of Minnesota. The tourism industry makes significant contributions to the state's economy, at the same time, relying on the many natural assets that the state has to continue attracting visitors. Therefore, it is important that the tourism industry contributes to, rather than deters, the progress of sustainability practices. We believe that releasing this dataset will help increase public awareness of and interest in the trend of implementing sustainability practices among tourism businesses in Minnesota.

Date: 2013-10-29

Coverage,
Temporal: 2007-2013

Provenance The same online survey was administered in April of 2007, 2010, and 2013. Each time the survey was administered, an SPSS data file was generated. The three SPSS data files were merged to create the current data file that includes all the data collected in 2007, 2010, and 2013.

Grants: Not applicable

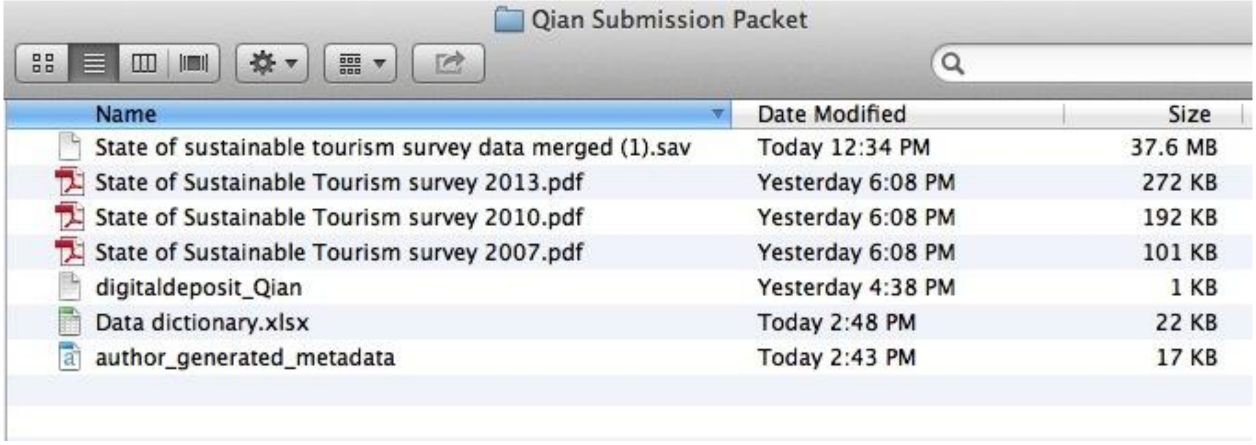
Additional Information

On submission, Qian shared some information about the dataset with you via phone:

- These data are the result of partnering with the Explore MN, a tourism unit of the state, which provided the participant pool for the study (300-500 respondents).
- The SPSS file included variable names with labels indicating how the data were coded for analysis. Qian created a data dictionary to document this information as well as capturing which questions did not occur on all three surveys.
- Qian was unsure if the survey instruments were proprietary or not. But after verifying with the Tourism Center, found that she could release the questions with the dataset for public reuse. In additional, all identifying information for the participants have been removed.
- Qian feels that she owns the data and has signed a deposit agreement for the library.

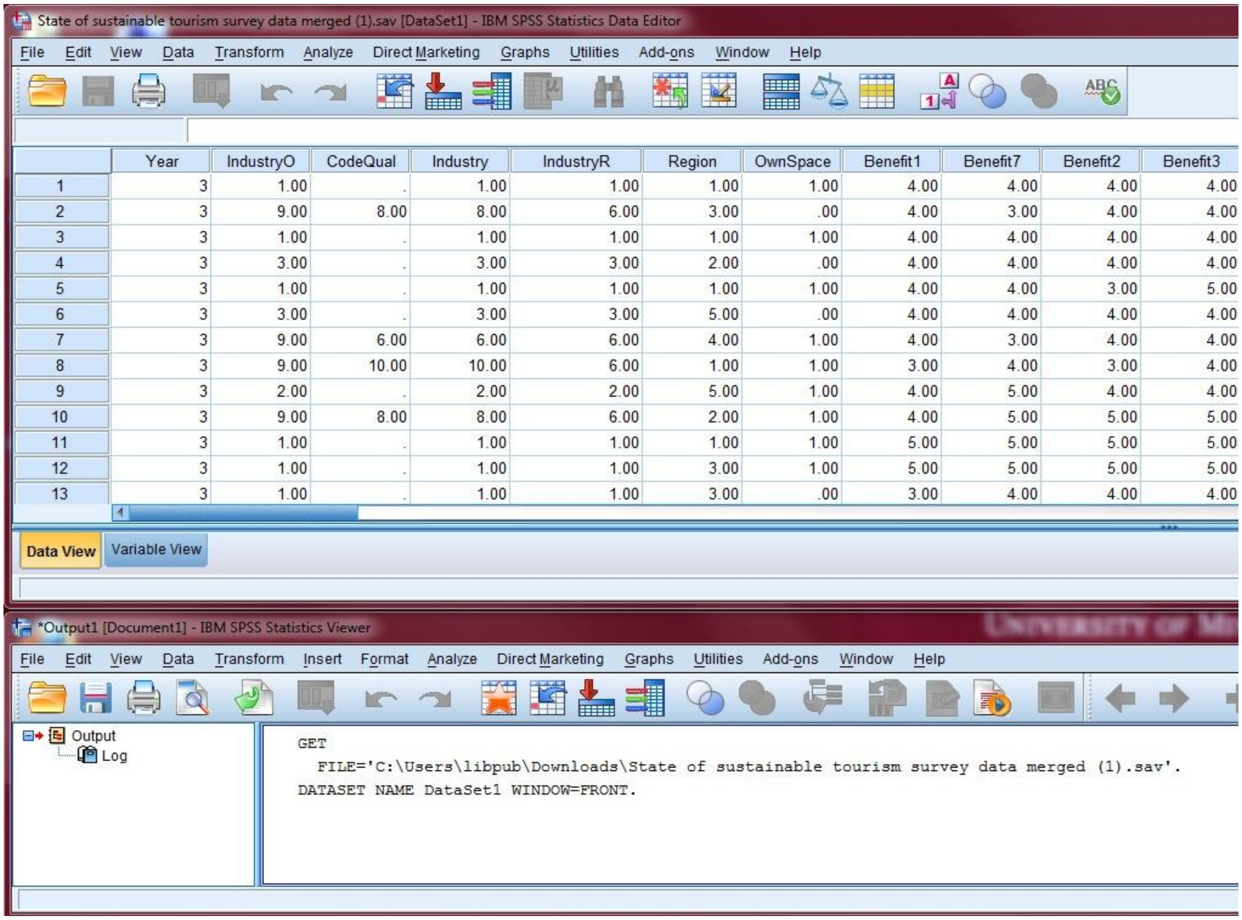
Submission Information Package (SIP):

Here are all the files received from Dr. Qian, including the metadata information that is written above, called author_generated_metadata.rtf.



Name	Date Modified	Size
State of sustainable tourism survey data merged (1).sav	Today 12:34 PM	37.6 MB
State of Sustainable Tourism survey 2013.pdf	Yesterday 6:08 PM	272 KB
State of Sustainable Tourism survey 2010.pdf	Yesterday 6:08 PM	192 KB
State of Sustainable Tourism survey 2007.pdf	Yesterday 6:08 PM	101 KB
digitaldeposit_Qian	Yesterday 4:38 PM	1 KB
Data dictionary.xlsx	Today 2:48 PM	22 KB
author_generated_metadata	Today 2:43 PM	17 KB

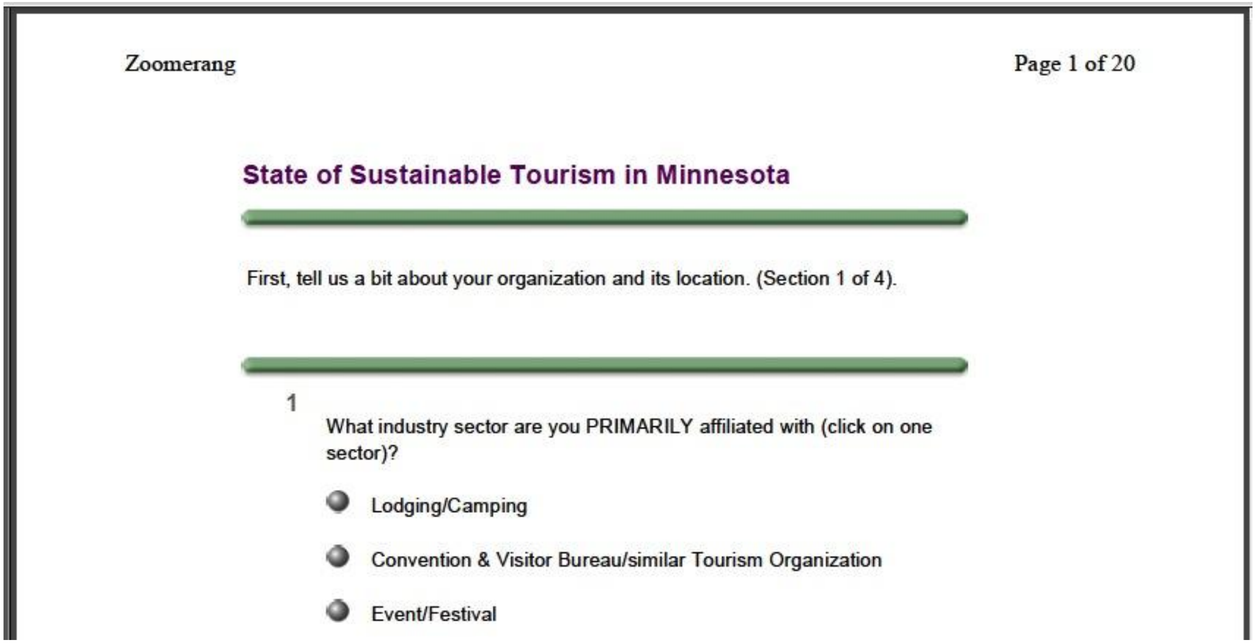
Screenshot of State of sustainable tourism survey data merged (1).sav



	Year	IndustryO	CodeQual	Industry	IndustryR	Region	OwnSpace	Benefit1	Benefit7	Benefit2	Benefit3
1	3	1.00	.	1.00	1.00	1.00	1.00	4.00	4.00	4.00	4.00
2	3	9.00	8.00	8.00	6.00	3.00	.00	4.00	3.00	4.00	4.00
3	3	1.00	.	1.00	1.00	1.00	1.00	4.00	4.00	4.00	4.00
4	3	3.00	.	3.00	3.00	2.00	.00	4.00	4.00	4.00	4.00
5	3	1.00	.	1.00	1.00	1.00	1.00	4.00	4.00	3.00	5.00
6	3	3.00	.	3.00	3.00	5.00	.00	4.00	4.00	4.00	4.00
7	3	9.00	6.00	6.00	6.00	4.00	1.00	4.00	3.00	4.00	4.00
8	3	9.00	10.00	10.00	6.00	1.00	1.00	3.00	4.00	3.00	4.00
9	3	2.00	.	2.00	2.00	5.00	1.00	4.00	5.00	4.00	4.00
10	3	9.00	8.00	8.00	6.00	2.00	1.00	4.00	5.00	5.00	5.00
11	3	1.00	.	1.00	1.00	1.00	1.00	5.00	5.00	5.00	5.00
12	3	1.00	.	1.00	1.00	3.00	1.00	5.00	5.00	5.00	5.00
13	3	1.00	.	1.00	1.00	3.00	.00	3.00	4.00	4.00	4.00

```
GET
  FILE='C:\Users\libpub\Downloads\State of sustainable tourism survey data merged (1).sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

Screenshot of State of Sustainable Tourism survey 2007.pdf



Zoomerang

Page 1 of 20

State of Sustainable Tourism in Minnesota

First, tell us a bit about your organization and its location. (Section 1 of 4).

1

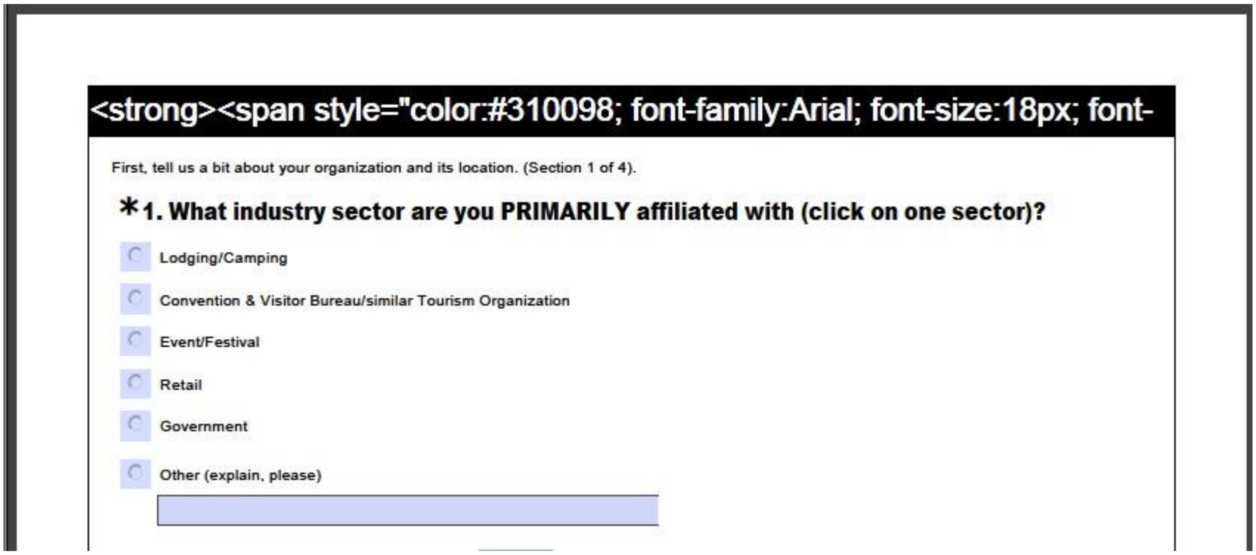
What industry sector are you PRIMARILY affiliated with (click on one sector)?

● Lodging/Camping

● Convention & Visitor Bureau/similar Tourism Organization

● Event/Festival

Screenshot of State of Sustainable Tourism survey 2010.pdf



<strong><span style="color:#310098; font-family:Arial; font-size:18px; font-

First, tell us a bit about your organization and its location. (Section 1 of 4).

*1. What industry sector are you PRIMARILY affiliated with (click on one sector)?

○ Lodging/Camping

○ Convention & Visitor Bureau/similar Tourism Organization

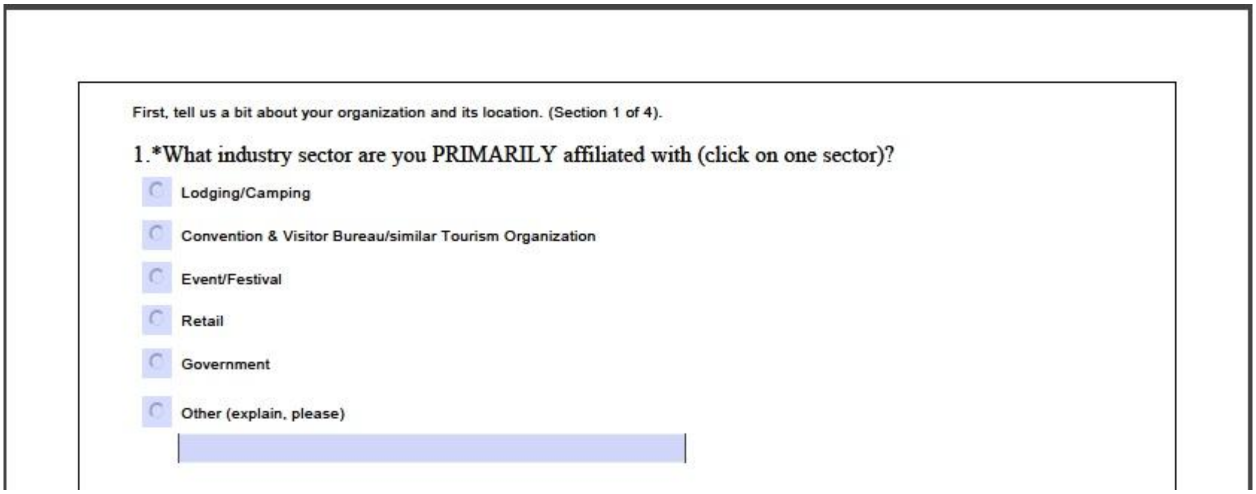
○ Event/Festival

○ Retail

○ Government

○ Other (explain, please)

Screenshot of State of Sustainable Tourism survey 2013.pdf



First, tell us a bit about your organization and its location. (Section 1 of 4).

1.*What industry sector are you PRIMARILY affiliated with (click on one sector)?

○ Lodging/Camping

○ Convention & Visitor Bureau/similar Tourism Organization

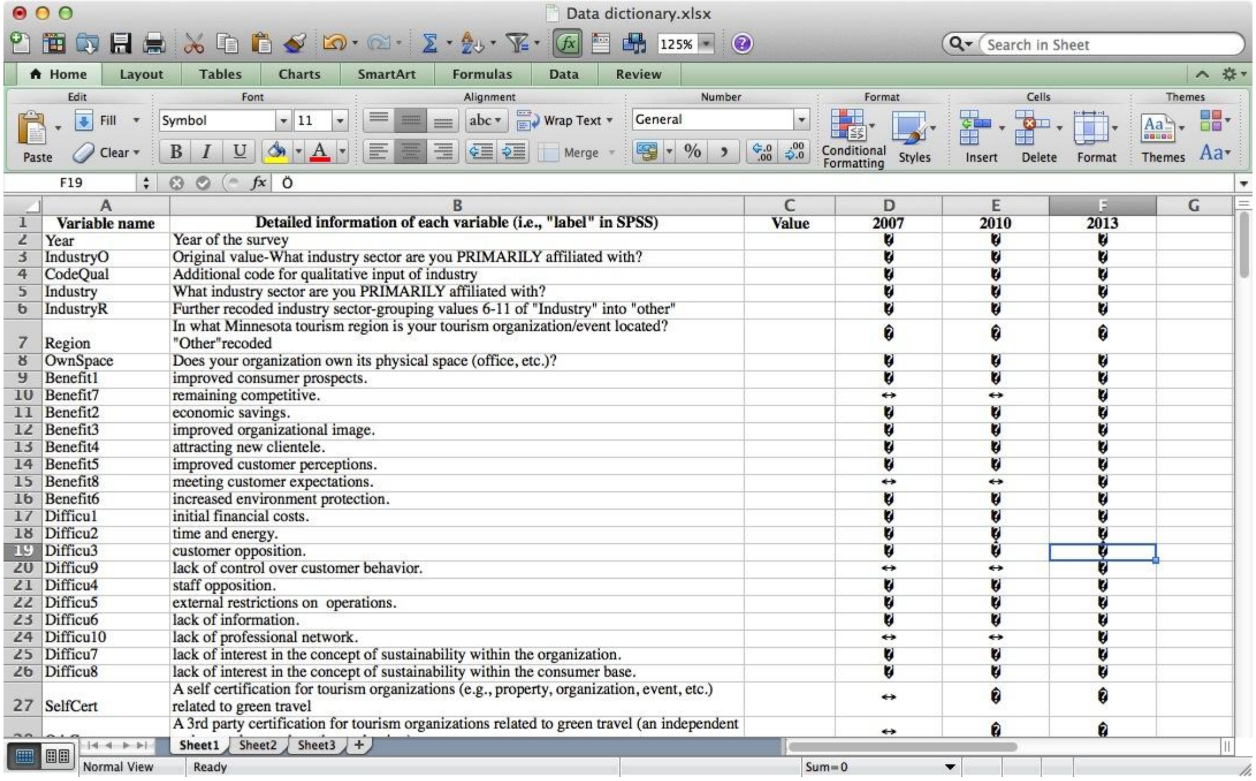
○ Event/Festival

○ Retail

○ Government

○ Other (explain, please)

Screenshot of Data dictionary.xls



F19 fx Ö

	A	B	C	D	E	F	G
1	**Variable name**	**Detailed information of each variable (i.e., "label" in SPSS)**	**Value**	**2007**	**2010**	**2013**	
2	Year	Year of the survey		✓	✓	✓	
3	IndustryO	Original value-What industry sector are you PRIMARILY affiliated with?		✓	✓	✓	
4	CodeQual	Additional code for qualitative input of industry		✓	✓	✓	
5	Industry	What industry sector are you PRIMARILY affiliated with?		✓	✓	✓	
6	IndustryR	Further recoded industry sector-grouping values 6-11 of "Industry" into "other"		✓	✓	✓	
7	Region	In what Minnesota tourism region is your tourism organization/event located? "Other"recoded		?	?	?	
8	OwnSpace	Does your organization own its physical space (office, etc.)?		✓	✓	✓	
9	Benefit1	improved consumer prospects.		✓	✓	✓	
10	Benefit7	remaining competitive.		↔	↔	✓	
11	Benefit2	economic savings.		✓	✓	✓	
12	Benefit3	improved organizational image.		✓	✓	✓	
13	Benefit4	attracting new clientele.		✓	✓	✓	
14	Benefit5	improved customer perceptions.		✓	✓	✓	
15	Benefit8	meeting customer expectations.		↔	↔	✓	
16	Benefit6	increased environment protection.		✓	✓	✓	
17	Difficu1	initial financial costs.		✓	✓	✓	
18	Difficu2	time and energy.		✓	✓	✓	
19	Difficu3	customer opposition.		✓	✓	✓	
20	Difficu9	lack of control over customer behavior.		↔	↔	✓	
21	Difficu4	staff opposition.		✓	✓	✓	
22	Difficu5	external restrictions on operations.		✓	✓	✓	
23	Difficu6	lack of information.		✓	✓	✓	
24	Difficu10	lack of professional network.		↔	↔	✓	
25	Difficu7	lack of interest in the concept of sustainability within the organization.		✓	✓	✓	
26	Difficu8	lack of interest in the concept of sustainability within the consumer base.		✓	✓	✓	
27	SelfCert	A self certification for tourism organizations (e.g., property, organization, event, etc.) related to green travel		↔	?	?	
		A 3rd party certification for tourism organizations related to green travel (an independent		↔	?	?	

C.5 Social Sciences/Humanities Data

Data Curation Pilot 2013: Humanities/Social Sciences Data

Dr. Mary Hermes is an education researcher in the Curriculum and Instruction Department. Her primary research interest is developing multimedia tools to share Ojibwe language and culture.

Type of Data

Dr. Hermes submitted the following dataset to be curated for reuse in the library. These data are a video (3.34m), an audio file, and a transcript of the movie that records native Ojibwe speakers engaged in everyday activities. The video has been transcribed using ELAN software to track the words and timestamps of the audio track. Then each transcription is translated from Ojibwe to English in a MS word document. These videos are of great research value as most language research is done using a standard question/response approach, whereas Hermes is capturing discussion around everyday activities, such as making bread or doing the laundry. (Note: Hermes has about 25 sets of these data, each a different movie, of which one has been selected for the pilot. The processed video files can be large, from 586MB up to 85GB.)

Metadata

Hermes, on submission, has not filled out any metadata fields.

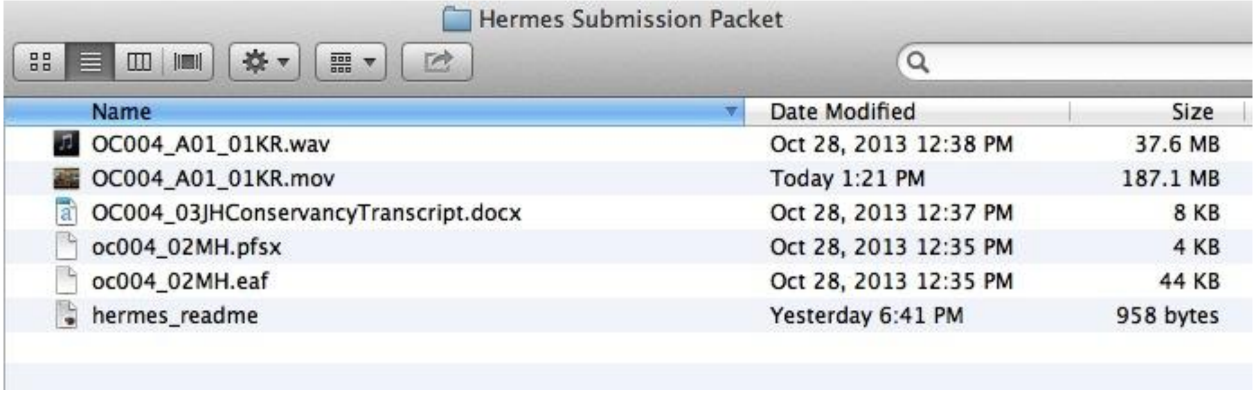
Additional Information

On submission, Hermes shared some information about the dataset with you via phone:

- The data do not have any confidential material and all participants were informed that these would be shared for research use.
- The transcriptions would be best presented in a format that would allow students and researchers the ability to re-transcript or augment.
- The common file format for transcriptions is EAF (ELAN Annotation Format) which is an XML-based, documented standard for endangered languages.
- To open .eaf files (or .pfsx in Windows), anyone can download the open source ELAN tool, for the creation of complex annotations on video and audio resources (http://tla.mpi.nl/tools/tla-tools/elan/).
- This is an NSF funded grant project that has already received a lot of press.

Submission Information Package (SIP):

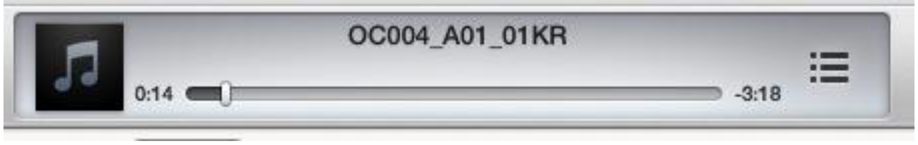
Here are all the files received from Dr. Hermes, including a note from her graduate assistant explaining that an audio file and the metadata information is not yet ready (hermes_readme.txt).



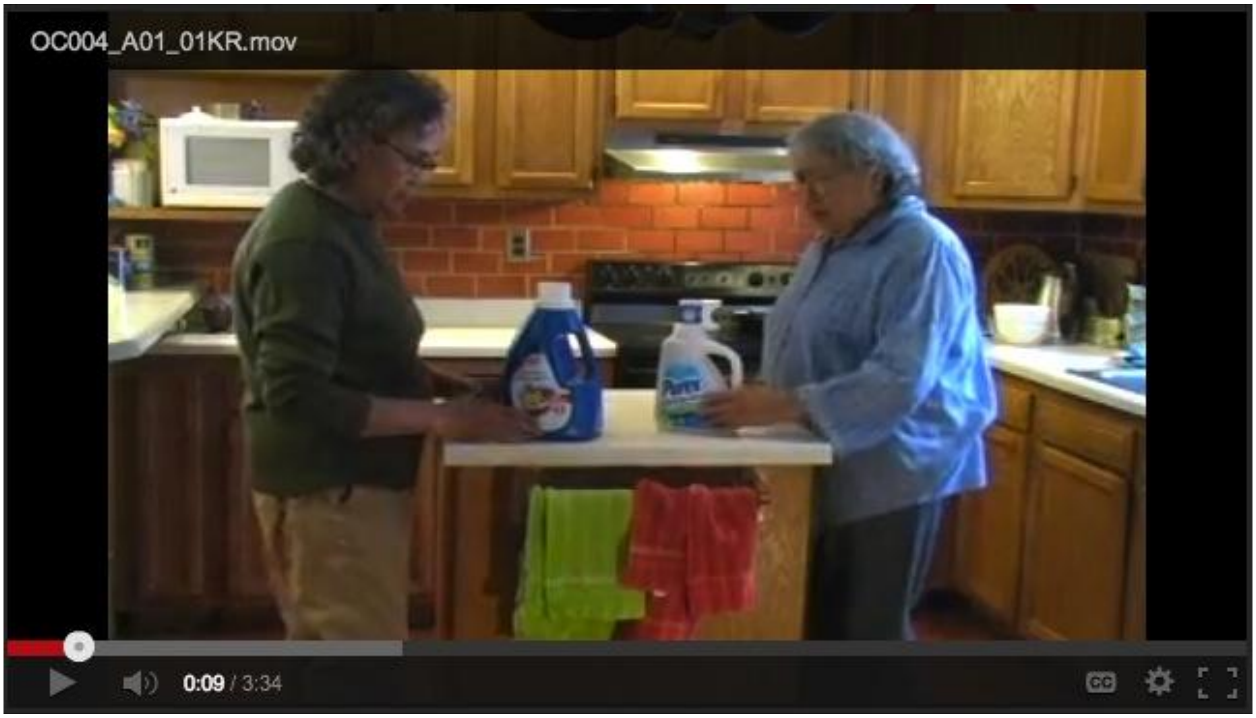
Hermes Submission Packet

Name	Date Modified	Size
OC004_A01_01KR.wav	Oct 28, 2013 12:38 PM	37.6 MB
OC004_A01_01KR.mov	Today 1:21 PM	187.1 MB
OC004_03JHConservancyTranscript.docx	Oct 28, 2013 12:37 PM	8 KB
oc004_02MH.pfsx	Oct 28, 2013 12:35 PM	4 KB
oc004_02MH.eaf	Oct 28, 2013 12:35 PM	44 KB
hermes_readme	Yesterday 6:41 PM	958 bytes

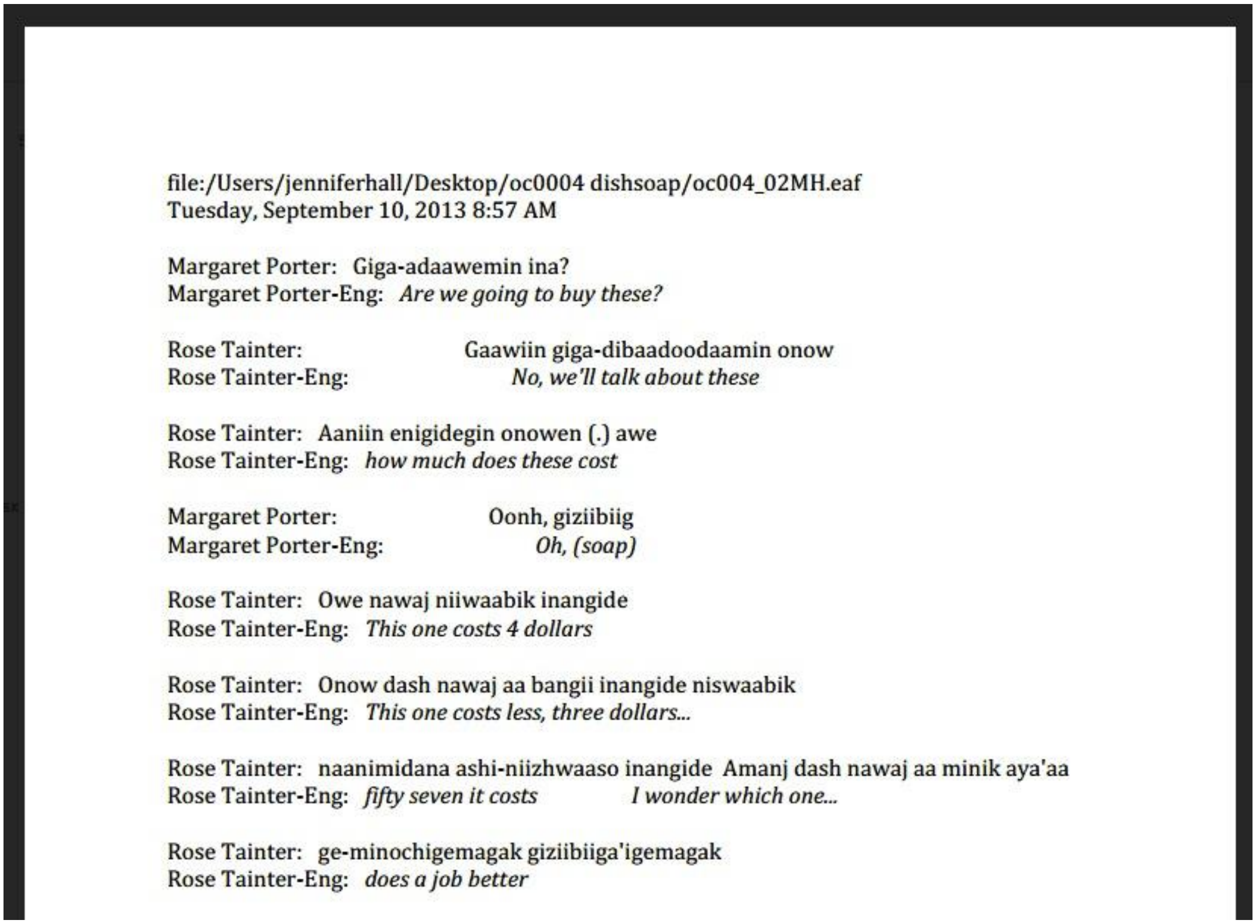
Screenshot of OC004_A01_01KR.wav



Screenshot of OC004_A01_01KR.mov



Screenshot of OC004_03JHConservancyTranscript.doc



file:/Users/jenniferhall/Desktop/oc0004 dishsoap/oc004_02MH.eaf
Tuesday, September 10, 2013 8:57 AM

Margaret Porter: Giga-adaawemin ina?
Margaret Porter-Eng: *Are we going to buy these?*

Rose Tainter: Gaawiin giga-dibaadoodaamin onow
Rose Tainter-Eng: *No, we'll talk about these*

Rose Tainter: Aaniin enigidegin onowen (.) awe
Rose Tainter-Eng: *how much does these cost*

Margaret Porter: Oonh, giziibiig
Margaret Porter-Eng: *Oh, (soap)*

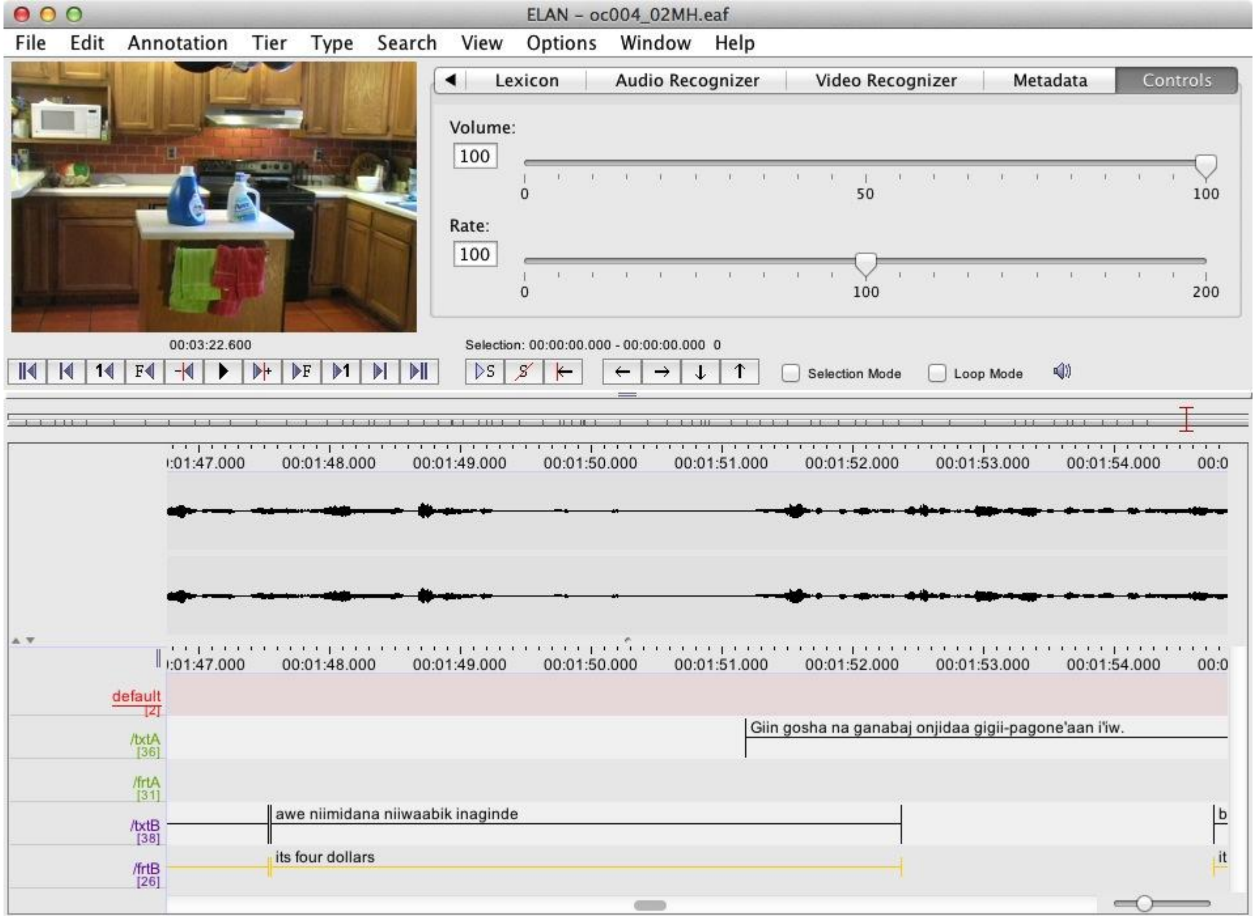
Rose Tainter: Owe nawaj niiwaabik inangide
Rose Tainter-Eng: *This one costs 4 dollars*

Rose Tainter: Onow dash nawaj aa bangii inangide niswaabik
Rose Tainter-Eng: *This one costs less, three dollars...*

Rose Tainter: naanimidana ashi-niizhwaaso inangide Amanj dash nawaj aa minik aya'aa
Rose Tainter-Eng: *fifty seven it costs* *I wonder which one...*

Rose Tainter: ge-minochigemagak giziibiiga'igemagak
Rose Tainter-Eng: *does a job better*

Screenshot of oc004_02MH.eaf (using ELAN software for MAC)

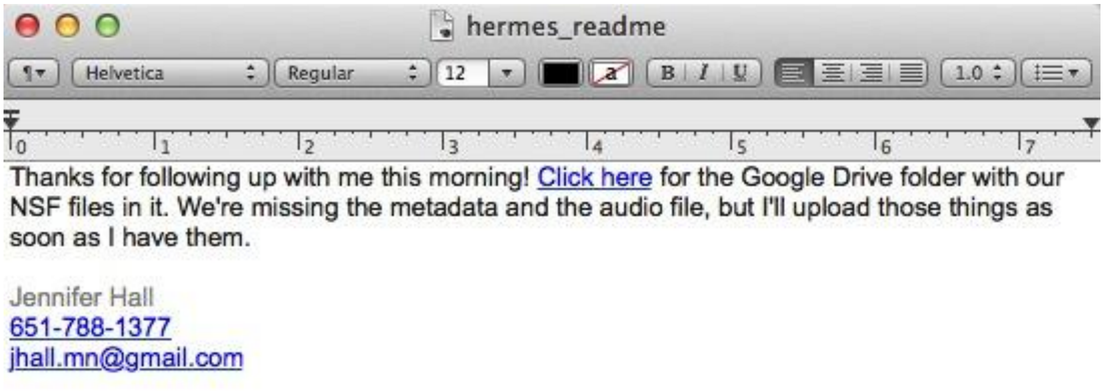


Giin gosha na ganabaj onjidaa gigii-pagone'aan i'iw.

awe niimidana niiwaabik inaginde

its four dollars

Screenshot of hermes_readme.txt



Thanks for following up with me this morning! Click here for the Google Drive folder with our NSF files in it. We're missing the metadata and the audio file, but I'll upload those things as soon as I have them.

Jennifer Hall
651-788-1377
jhall.mn@gmail.com

APPENDIX D: AGENDA FOR THE NOVEMBER 4TH DATA CURATION PILOT: DIGITAL SANDBOX

Digital Curation Sandbox: Data Pilot 2013

Where: STSS 512 B
When: November 4th 2013 9am-12pm

Overview
The Libraries' 2013 Data Curation Pilot project (http://z.umn.edu/datapilot13) has accepted five datasets to be curated for public use using the libraries existing infrastructure. This will help us better understand the data access, archiving, and preservation needs of our campus users. To do this, we need your help!

This hands-on event will team a subject librarian, a digital technologist, and an archivist/curator to appraise one of the digital datasets received in our pilot and determine a treatment process for that data. The team will include a facilitator that will help lead the conceptual process and discussion. Then, we will compare results and begin to develop some common data curation workflows that span our disciplinary data examples. Refreshments will be served.

Objectives of the Event:

- To create a treatment process for each of the 5 datasets received in the 2013 Data Curation Project that utilizes the skills and best practices of library staff, including archivists, digital technologies, and subject librarians.
- To complete an analysis of the 5 workflows that identifies common elements, which may become the foundation of a more generalizable treatment process or curation workflow for research data.

Objectives for participants:

- To gain/increase hands-on experience with a digital curation workflow process in order to directly apply them to digital research data.
- To become/increase familiarity with the problems and difficulties that arise with data curation in order to better assess and facilitate future directions of the libraries in this area.

Instructions for Event
Seat the group into 5 teams with a subject liaison, an archivist/curator, and a digital technologist (28-30 participants). Each team will also include a "data" facilitator who is prepped on the data curation conceptual model and will help your group take notes.

Agenda

9:00 AM	Welcome (10 min) - Lisa • Introduce topic of digital curation • Outline agenda for the day Icebreaker for tables (5 min) • Introduce yourself to the table and describe your "role" and how it might support to digital curation.
9:15	Activity #1: Engage with the digital curation lifecycle using fun, easy example (25 min) • Introduce Dr. Watson's collection (analog) • Teams write steps to curate Dr Watsons' collection o Individual post-it notes o group them into categories to represent a workflow o Facilitator has blue post it notes to capture categories, Facilitator writes/posts major categories in order on board with arrows

	o Facilitator presents workflow
9:40	Data Curation Pilot Overview (5 mins) - Lisa • Learn from our experience to apply to digital. • Moreover, what if Watson had a federal mandate to share all of this data? • Introduce the 2013 Data Curation Pilot
9:45	Activity #2: Break into 5 groups of 4 and discuss your example dataset. Each group would be assigned one dataset from our 5 examples. (60 min) o Liaison introduces data and research. Discuss how the interview went. o Draft a treatment process for curation based on your data se ▪ List each curation stage in your spreadsheet/or up on the board ▪ For each stage list • Curation Stage (will, etc.) • Activities for each stage • Questions to consider for each stage • What Actions should be take with this particular dataset.
10:45	Break (10 min)
10:55	Report and questions for each group: Share the 5 treatments with entire group. (30 min)
11:25	Activity #3: As a large group, discuss/compare processes and define the common elements that might be a baseline process for curating datasets in the library. (30 min)
11:55	Wrap up and next steps (5 min) - Lisa
12:00 PM	End of Session (Total 180 min with 5 min flex)

Seating Chart (by table)

Engineering Data: David Levinson, Civil Engineering / Road Maps and GIS

- Jon Jeffryes
- Stephen Hearn
- Christine DeZelar-Tiedman
- Facilitator: Ryan Mattke

Health Sciences Data: James Hodges, School of Public Health / Clinical Trial Excel Data

- Steven Braun
- Jon Nichols
- Erik Moore
- Facilitator: Meghan Lafferty

Humanities/Social Sciences Data: Mary Hermes: Curriculum and Instruction / Ojibwe Video and Transcriptions

- Kim Clarke
- Jason Roy
- Lois Hendrickson
- Facilitator: Justin Schell

Natural Resources Data: Lisa Qian, Extension (Tourism Studies) / Survey and SPSS Data

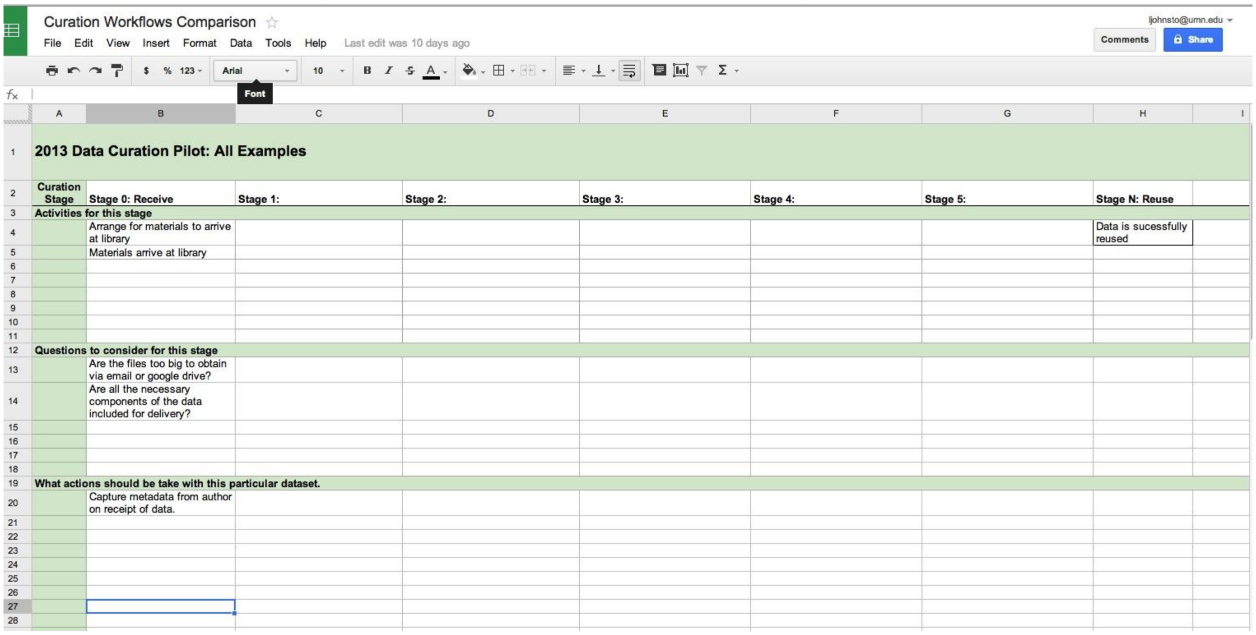
- Kristen Mastel / Shannon Farrell
- Carol Kussmann

- Lara Friedman-Shedlov
- Facilitator: Amy West / Alicia Hofelich Mohr

Interdisciplinary Data: Bill Arnold / Longitudinal Excel Data

- Josh Bishoff
- Bill Tantzen / John Butler
- Daniel Necas
- Facilitator: Carolyn Rauber / Francine Dupont Crocker

APPENDIX E: CURATION WORKFLOW WORKSHEET FOR PILOT DATA



	A	B	C	D	E	F	G	H	I
1	2013 Data Curation Pilot: All Examples								
2	Curation Stage	Stage 0: Receive	Stage 1:	Stage 2:	Stage 3:	Stage 4:	Stage 5:	Stage N: Reuse	
3	Activities for this stage								
4		Arrange for materials to arrive at library						Data is sucessfully reused	
5		Materials arrive at library							
6									
7									
8									
9									
10									
11									
12	Questions to consider for this stage								
13		Are the files too big to obtain via email or google drive?							
14		Are all the necessary components of the data included for delivery?							
15									
16									
17									
18									
19	What actions should be take with this particular dataset.								
20		Capture metadata from author on receipt of data.							
21									
22									
23									
24									
25									
26									
27									
28									