

**Syllabus: LIS 855, Digital curation**  
School of Library and Information Studies  
University of Wisconsin-Madison  
Spring 2010: Online

Dorothea Salo (please call me "Dorothea")  
Office address: 330B Memorial Library  
AIM: mindsatuw (work), DorotheaSalo (home)  
Course link page: <http://pinboard.in/u:dsalo/t:855>

dsalo@library.wisc.edu, 262-5493  
Office Hours: 7-9 pm Tuesdays  
Skype: dorotheasalo

## Course Objectives

- Ability to evaluate and work with digital-preservation software tools.
- Ability to read and write XML metadata, particularly METS, MODS, and PREMIS.
- Ability to assess, plan for, manage, and execute a small-scale digital-preservation project.
- Ability to assess digital data for preservability; ability to make yes-or-no accessioning decisions.
- Understanding of technological, economic, and social models of digital preservation.
- Understanding of data forms and formats, and data lifecycles in different scholarly disciplines.
- Construction of a current-awareness strategy; ability to assimilate substantial amounts of relevant writing.
- Sufficient courage, self-awareness, and skill for self-sufficiency in acquiring technical knowledge.

## Course Policies

**\*\*It is my desire to fully include persons with disabilities in this course. Please let me know within two weeks if you require special accommodation. I will try to maintain the confidentiality of this information.\*\***

Academic Honesty: I follow the academic standards for cheating and plagiarism set forth by the University of Wisconsin.

An explicit goal of this course is self-sufficiency in acquiring knowledge about novel technology. To that end, I will NOT go into exhaustive detail on every technology we look at. You are expected to exhaust normal information channels before you approach classmates or (especially) me with nuts-and-bolts technology questions.

## Grading Schema and Due Dates

<u>Assignments:</u>	<u>Percentage</u>	<u>Due Date</u>
In-class assignments	50%	(various)
Final project	50%	9 May

There are no extra credit opportunities available in this class. No assignment grades will be dropped; any student failing to turn in a major assignment will automatically fail the course. Particularly with technical assignments, perfection is not the goal; learning is. Failures and retrenchments are to be expected, and usually will not count against your grade.

## Unit 1: The sociology of digital data management and preservation

### Week of January 17: Course overview and XML introduction

*Learning objectives: XML well-formedness rules. DTDs, schemas, XML validity. XML editors. Declaring and using XML namespaces.*

*Assignment (due 1/24): XML well-formedness.*

Gold, Anna. "Cyberinfrastructure, data, and libraries." D-Lib Magazine 13:9/10 (2007). <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html> and <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>

ARL, "Agenda for developing e-science." [http://www.arl.org/bm~doc/ARL\\_EScience\\_final.pdf](http://www.arl.org/bm~doc/ARL_EScience_final.pdf) (pp. 3-13)

w3schools.com. "XML tutorial." <http://www.w3schools.com/xml/default.asp> (Introduction, How to Use, Syntax, Elements, Attributes, then below to XML Namespaces and XML Summary)

"Comparison of XML editors." [http://en.wikipedia.org/wiki/List\\_of\\_XML\\_editors](http://en.wikipedia.org/wiki/List_of_XML_editors)

### Week of January 24: Types of data. Data lifecycle models. Data interviews.

*Learning objectives: OAIS model. DCC data-lifecycle model. Types and sources of scientific data. Reference, research, and resource data collections. Examples of quantitative and qualitative social-science data. Examples and uses of humanities data. Crowdsourcing data transcription and analysis. Data curation profiles. Data interviews.*

*Assignment (part of project plan): Schedule and perform data interview with final-project client.*

- Ockerbloom, John Mark. "What repositories do: the OAIS model." <http://everybodyslibraries.com/2008/10/13/what-repositories-do-the-oais-model/>
- OAIS Reference Model. <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- DCC Curation Lifecycle Model. <http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf>
- Lifecycle Model FAQs <http://www.dcc.ac.uk/resources/curation-lifecycle-model/lifecycle-model-faqs>
- ICPSR. "Guide to social science data preparation and archiving." <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>
- Cragin, Melissa, and Kalpana Shankar. "Scientific data collections and distributed collective practice." *Computer Supported Cooperative Work* 15:2/3 (2006). <http://dx.doi.org/10.1007/s10606-006-9018-z>
- Raloff, Janet. "Galaxy Zoo's blue mystery." *ScienceNews*. [http://www.sciencenews.org/view/generic/id/33403/title/Science\\_%2B\\_the\\_Public\\_\\_Galaxy\\_Zoos\\_blue\\_mystery\\_%28part\\_I%29](http://www.sciencenews.org/view/generic/id/33403/title/Science_%2B_the_Public__Galaxy_Zoos_blue_mystery_%28part_I%29) and [http://www.sciencenews.org/view/generic/id/33436/title/Galaxy\\_Zoos\\_blue\\_mystery\\_%28part\\_2%29](http://www.sciencenews.org/view/generic/id/33436/title/Galaxy_Zoos_blue_mystery_%28part_2%29)
- Mueller, Martin. "Getting undergraduates and amateurs into the business of re-editing our cultural heritage." <http://literaryinformatics.net/2011/01/07/getting-undergraduates-and-amateurs-into-the-business-of-re-editing-our-cultural-heritage-for-a-digital-world/>
- Witt, Michael, and Jake R. Carlson. "Conducting a data interview." [http://docs.lib.purdue.edu/lib\\_research/81/](http://docs.lib.purdue.edu/lib_research/81/)
- Data Curation Profiles Toolkit. <http://www4.lib.purdue.edu/dcp/> (Please register for the site, download and read all the materials linked from <http://www4.lib.purdue.edu/dcp/download>, and read at least two "Completed Profiles.")

### **Week of January 31: Project management**

*Learning objectives: Project management techniques. Project planning. Dealing with stakeholders. Critical path analysis. Budgeting and cost estimates. Monitoring progress. Running meetings. Common pitfalls.*

*Assignment (due 2/14): Project plan for final project.*

- German, Lisa. "No one plans to fail; they fail to plan." *Technicalities* 29:3 (2009). <http://vnweb.hwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e1186e01681fd1ff70cc611d6bbae89c2032582a6bf649d8a613bee0b00989a5a&fmt=H>
- Marill, Jennifer L., and Marcella Leshner. "Mile high to ground level." *The Serials Librarian* 52:3 (2007). [http://dx.doi.org/10.1300/J123v52n03\\_11](http://dx.doi.org/10.1300/J123v52n03_11)
- Wamsley, Lori H. "Controlling project chaos: project management for library staff." *PNLA Quarterly* 73:2 (2009). [http://www.pnla.org/quarterly/Winter2009/PNLA\\_Winter09.pdf](http://www.pnla.org/quarterly/Winter2009/PNLA_Winter09.pdf) (pp. 5-6, 27)
- Cervone, H. Frank. "Good project managers are clueful rather than clueless." *OCLC Systems and Services* 24:4 (2008). <http://dx.doi.org/10.1108/10650750810914201>
- Read through the DoIT Project Management Advisor at <http://pma.doit.wisc.edu/>

### **Week of February 7: Researcher practices and needs**

*Learning objectives: Scholarly attitudes toward data sharing, and how they differ across disciplines. Personal digital preservation, and how practices bleed into the scholarly environment. Researcher attitudes toward librarians and archivists. Data security. Data citation and credit (Datacite, ORCID). Journals and data. Peer review and data.*

- Marshall, Catherine C. "Rethinking personal digital archiving." <http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html> and <http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html>
- Borgman, Christine L. "Research data: who will share what, with whom, when, and why?" <http://works.bepress.com/borgman/238/>
- Brown, C. Titus. "My data management plan -- a satire." <http://ivory.idyll.org/blog/may-10/data-management.html>
- Wilson, James A.J. and Meriel Patrick. "Sudamih researcher requirements report." <http://sudamih.oucs.ox.ac.uk/docs/Sudamih%20Researcher%20Requirements%20Report.pdf>
- Lawrence, Bryan. "Citation, digital object identifiers, persistence, correction, and metadata." [http://home.badc.rl.ac.uk/lawrence/blog/2011/01/07/citation,\\_digital\\_object\\_identifiers,\\_persistence,\\_correction\\_and\\_metadata](http://home.badc.rl.ac.uk/lawrence/blog/2011/01/07/citation,_digital_object_identifiers,_persistence,_correction_and_metadata)
- Timmer, John "Jaz drives, spiral notebooks, and SCSI: how we lose scientific data." <http://arstechnica.com/science/news/2010/11/preserving-science-how-data-gets-lost.ars>
- Karger, David. "Why all your data should live in one application." <http://groups.csail.mit.edu/haystack/blog/2010/10/20/why-all-your-data-should-live-in-one-application/>
- "What is DataCite?" <http://datacite.org/whatisdc.html>
- Fenner, Martin. "ORCID or how to build a unique identifier for scientists in 10 easy steps." <http://blogs.nature.com/mfenner/2010/01/03/orcid-or-how-to-build-a-unique-identifier-for-scientists-in-10-easy-steps>

## **Week of February 14: Sustainability and economic models**

*Learning objectives: Macro-economics of digital preservation. Perils of grant funding. Perils of governmental funding. Perils of institutional funding.*

*Assignment (due 2/21): Read archives of designated blogs; link-and-summarize one useful post.*

"Sustainable Economics for a Digital Planet." [http://brtf.sdsc.edu/biblio/BRTF\\_Final\\_Report.pdf](http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf)

Ithaka. "Funding sustainable digital resources." <http://www.ithaka.org/ithaka-s-r/research/funding-sustainable-digital-resources>

Goldstein, Serge J., and Mark Ratliff. "DataSpace: a funding and operational model." <http://arks.princeton.edu/ark:/88435/dsp01w6634361k>

Wilson et al. "Developing infrastructure for research data management at the University of Oxford." *Ariadne* 65 (2010). <http://www.ariadne.ac.uk/issue65/wilson-et-al/>

Timmer, John. "How science funding is putting scientific data at risk." <http://arstechnica.com/science/news/2010/10/how-science-funding-is-putting-scientific-data-at-risk.ars>

## **Week of February 21: The legal and regulatory environment**

*Learning objectives: Open movements (open source, open access, open data, open government data, open notebook science).*

*Funder mandates (NIH Public Access Policy, NSF DMPs). Journal open-data mandates. Copyright and data. Patents and data.*

*Panton Principles, CC0. The dangers of "non-commercial" and "share-alike" licenses. Human-subjects research and data confidentiality. Documenting traditional cultural expressions.*

*Assignment (due 2/28): Evaluate and suggest improvements to an NSF data-management plan.*

Salo, Dorothea. "Battle of the opens." <http://scientopia.org/blogs/bookoftrogool/2010/03/15/battle-of-the-opens/>

NIH. "Frequently asked questions about the NIH Public Access Policy." <http://publicaccess.nih.gov/FAQ.htm>

NSF. "Dissemination and sharing of research results." [http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag\\_6.jsp#VID4](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)

NSF. "Dissemination and sharing of research results." <http://www.nsf.gov/bfa/dias/policy/dmp.jsp> (Please skim all directorates' guidance. Pay special attention to guidance in any area where you have disciplinary expertise.)

"Share the data: making large-scale proteomics data widely available." *Bio-IT World*. <http://www.bio-itworld.com/2010/08/25/open-proteomics-comment.html>

Nguyen, Think. "Remembering Babel: open data sharing & integration." <http://sciencecommons.org/weblog/archives/2009/11/19/remembering-babel-open-data-sharing-integration/>

Nguyen, Think. "Freedom to research: keeping scientific data open, accessible, and interoperable." <http://sciencecommons.org/wp-content/uploads/freedom-to-research.pdf>

Panton Principles. <http://pantonprinciples.org/> and <http://pantonprinciples.org/faq/>

Bivens-Tatum, Wayne. "Librarians and traditional cultural expressions." <http://dx.doi.org/10.1080/10477845.2010.508693>

## **Week of February 28: Assessing data. Collection-development and digital-preservation policies.**

*Learning objectives: Data assessment. Gauging importance and preservability. Collection-development policies.*

*Assignments (due 3/7): Suggest/select topics for free week. Project groups: check in with instructor.*

Timmer, John. "Preserving science: what data do we keep?" <http://arstechnica.com/science/news/2010/11/preserving-science-choosing-what-data-to-discard.ars>

Skinner and Schultz, "Preserving Our Collections, Preserving Our Missions." [http://www.metaarchive.org/sites/default/files/GDDP\\_Educopia.pdf](http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf) (pp. 1-9)

Whyte, Angus. "Appraise & select research data for curation." <http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data>

Faundeen, John L., and Lyndon R. Oleson. "Scientific data appraisals: the value driver for preservation." [http://www.pv2007.dlr.de/Papers/Faundeen\\_AppraisalsValue\\_for\\_Preservation.pdf](http://www.pv2007.dlr.de/Papers/Faundeen_AppraisalsValue_for_Preservation.pdf)

Frigoli, Julia, Anne M. Etgen, and Michael Kuhar. "Developing and communicating responsible data management policies to trainees and colleagues." *Science and Engineering Ethics* (2010). <http://dx.doi.org/10.1007/s11948-010-9203-9>

Beagrie et al. "Digital preservation policies study." [http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy\\_p1finalreport.pdf](http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf)

## **Week of March 7: Library and archive preparedness**

*Learning objectives: Staffing models in libraries and archives. Job opportunities in data curation and digital preservation.*

*Embedded librarianship. Starting a brand-new data-curation program. Digital preservation needs and strategies in public*

libraries. Infrastructure. Funding (grant earmarks, budget and position shifting). Liaison librarians and research-data curation. Research data as “the new special collections.”

Assignment (due 3/28): Outlining a new data-curation program.

Swan, Alma. “Skills, role & career structure of data scientists & curators.” <http://www.jisc.ac.uk/publications/reports/2008/dataskillscareersfinalreport.aspx>

Rusbridge, Chris. “Tomorrow, and tomorrow, and tomorrow: poor players on the digital curation stage.” <http://www.era.lib.ed.ac.uk/handle/1842/2150/>

Pryor, Graham. “Librarians doing data -- a paradox?” [http://www.dcc.ac.uk/webfm\\_send/319](http://www.dcc.ac.uk/webfm_send/319)

Meyer, Lars. “Safeguarding collections.” <http://www.arl.org/bm~doc/safeguarding-collections.pdf>

Newton, Mark P., C. C. Miller, and Marianne Stowell Bracke. “Librarian roles in institutional repository data set collecting.” *Collection Management* 36:1 (2011). <http://dx.doi.org/10.1080/01462679.2011.530546>

Salo, Dorothea. “Retooling libraries for the data challenge.” *Ariadne* 64 (2010). <http://www.ariadne.ac.uk/issue64/salo/>

Westra, Brian. “Data services for the sciences: a needs assessment.” *Ariadne* 64 (2010). <http://www.ariadne.ac.uk/issue64/westra/>

**Week of March 14: SPRING BREAK**

**Week of March 21: STUDENT CHOICE and/or catchup week**

## Unit 2: The technology of digital preservation

**Week of March 28: Digitization, file formats, and digital sustainability**

*Learning objectives: Evaluating file formats for preservation. Lossy vs. lossless formats. Open vs. proprietary formats. File formats in instrument science. Quantitative-science file formats and tools (SPSS, Stata, R, Matlab). Image formats (JPEG, TIFF, JPEG 2000, PNG, GIF, RAW). Audio and video formats (codecs, sampling rate/bitrate, WAV, AIFF, mp3, MPEG4). HDF. GIS formats. “Preservation copy,” “digital surrogate.” Compound objects; archiving websites; BagIt.*

*Assignments: Crawling a static website for archival. Adding content files to METS file.*

Timmer, John. “Changing software, hardware a nightmare for tracking scientific data.” <http://arstechnica.com/science/news/2010/11/changing-software-hardware-a-nightmare-for-tracking-scientific-data.ars>

ICPSR, “Digital Preservation Tutorial,” section 3 “Obsolescence”: “File Formats and Software” and “Hardware and media” [http://www.icpsr.umich.edu/dpm/dpm-eng/eng\\_index.html](http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html)

Cornell, “Digital Imaging Tutorial.” <http://www.library.cornell.edu/preservation/tutorial/contents.html> (Skim.)

Rutgers, Video Object Standards Analysis, [http://rucore.libraries.rutgers.edu/collab/ref/dos\\_avwg\\_video\\_obj\\_standard.pdf](http://rucore.libraries.rutgers.edu/collab/ref/dos_avwg_video_obj_standard.pdf)

Rutgers, Audio Object Standards Analysis, [http://rucore.libraries.rutgers.edu/collab/ref/dos\\_avwg\\_audio\\_obj\\_standard.pdf](http://rucore.libraries.rutgers.edu/collab/ref/dos_avwg_audio_obj_standard.pdf)

Pilgrim, Mark. “A gentle introduction to video encoding.” <http://diveintomark.org/tag/give> (read all parts)

“Converting audio cassette tapes to CD, MP3, and other digital formats.” <http://www.andybrain.com/archive/convert-cassette-to-cd-digital.htm>

Farrell, Susan ed. “A guide to web preservation.” <http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf>

“Why HDF?” [http://www.hdfgroup.org/why\\_hdf/](http://www.hdfgroup.org/why_hdf/)

GIS DataDepot. “GIS data formats.” <http://data.geocomm.com/helpdesk/formats.html>

BagIt specification. <https://confluence.ucop.edu/display/Curation/BagIt> (please download and read the spec)

**Week of April 4: Metadata**

*Learning objectives: Descriptive, technical, administrative, and structural metadata. MODS, METS, PREMIS, Dublin Core. Discipline-specific metadata standards. Codebooks and data dictionaries. Crosswalking and crosswalking tools (Google Refine). Explaining metadata to non-librarians. Getting metadata from non-librarians. Coping with spreadsheets. DDI.*

*Assignment (due 4/18): Bare-bones METS file with MODS and PREMIS sections.*

Wilson, Andrew. “How much is enough: metadata for preserving digital data.” *Journal of Library Metadata* 10:2 (2010). <http://dx.doi.org/10.1080/19386389.2010.506395>

Riley, “Seeing Standards.” <http://www.dlib.indiana.edu/~jenlrile/metadatamap/> (Download the poster and read the legend and definitions carefully.)

Kennedy, “Nine questions to guide you in choosing a metadata schema.” <https://journals.tdl.org/jodi/article/viewArticle/226/205>

Guenther, McCallum, "New metadata standards for digital resources: MODS and METS." [http://findarticles.com/p/articles/mi\\_qa3991/is\\_200212/ai\\_n9150534](http://findarticles.com/p/articles/mi_qa3991/is_200212/ai_n9150534)

Cundiff and Trail, "Using METS and MODS..." <http://www.loc.gov/standards/mods/presentations/mets-mods-morgan-ala07/>

"Guidelines for using PREMIS with METS for exchange." <http://www.loc.gov/standards/premis/guidelines-premismets.pdf>

DDI FAQ. <http://www.ddialliance.org/resources/faq.html>

Getting started with DDI. <http://www.ddialliance.org/resources/getting-started>

### **Week of April 11: Threat models. Auditing. Organization. Versioning.**

*Learning objectives: Business-model and organizational threats to digital data. Risk assessment, analysis, and mitigation.*

*Migration and emulation, including tools. "Trusted digital repository." Dark archive. Two-tier archive. TRAC, DRAMBORA. File-format auditing tools (JHOVE, FITS). Bit-auditing and checksums.*

*Assignments (due 4/18): Using FITS; reading and evaluating its output; adding information to METS file. Project checkin.*

Ross, Seamus. "Preservation pressure points." <http://www.repositoryaudit.eu/images/PreservationPressurePoints.pdf>

Ross, Seamus, and Ann Gow. "Digital archaeology: rescuing neglected and damaged data resources." <http://eprints.erpanet.org/47/>

DRAMBORA Interactive. "DRAMBORA: About." <http://www.repositoryaudit.eu/about/> (Please register for the site and download the entire toolkit.)

CRL. "Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)." [http://www.crl.edu/sites/default/files/attachments/pages/trac\\_0.pdf](http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf)

CRL. "Report on Portico audit findings." <http://www.crl.edu/sites/default/files/attachments/pages/CRL%20Report%20on%20Portico%20Audit%202010.pdf>

JHOVE. <http://hul.harvard.edu/jhove/>

File Information Tool Set. <http://code.google.com/p/fits/wiki/tools>

### **Week of April 18: Hardware and software platforms for digital archival and preservation**

*Learning objectives: Hardware and its durability. Institutional repository platforms (DSpace, EPrints, Fedora, BePress Digital Commons, CONTENTdm). Digital-library platforms (Greenstone, ContentDM, Omeka). iRODS. Curation microservices. Relational databases, XML databases, RDF triplestores. Organizing files; choosing filenames. Versioning. Geographic dispersal techniques (LOCKSS, cloud storage, DuraCloud).*

*Assignment (due 5/2): Evaluating platforms for given content types.*

Murray, Peter. "Options in storage for digital preservation." <http://dltj.org/article/preservation-storage-options/> (following links strongly encouraged)

"About LOCKSS." [http://www.lockss.org/lockss/About\\_LOCKSS](http://www.lockss.org/lockss/About_LOCKSS)

"Top reasons to use DSpace." <http://www.dspace.org/why-use> (read skeptically!)

EPrints. <http://www.eprints.org/software/>

"Getting started with Fedora." <https://wiki.duraspace.org/display/FCR30/Getting+Started+with+Fedora>

"About Islandora." <http://islandora.ca/about> and [http://islandora.ca/solution\\_packs](http://islandora.ca/solution_packs)

"Advantages of Digital Commons." <http://www.bepress.com/ir/advantages.html>

"CONTENTdm overview." <http://www.oclc.org/contentdm/overview/default.htm>

"About Greenstone." <http://www.greenstone.org/>

"Omeka." <http://omeka.org/> (click around a bit)

"iRODS Overview." [https://www.irods.org/pubs/iRODS\\_Overview\\_0903.pdf](https://www.irods.org/pubs/iRODS_Overview_0903.pdf)

DuraCloud. "Introduction." <https://wiki.duraspace.org/display/duracloud/DuraCloud> (please read Features and Services also)

"Curation micro-services." <http://www.cdlib.org/services/uc3/curation/> (follow links, please)

"Merritt: An emergent micro-services approach to digital curation infrastructure." <https://confluence.ucop.edu/download/attachments/13860983/Merritt-latest.pdf>

Janée, Greg. "Resources, versions, and URIs." <http://www.alexandria.ucsb.edu/~gjane/archives/2009/versioning.html>

### **Week of April 25: E-records and records management**

*Learning objectives: Growth in e-records. Differences between e-records-management and other kinds of data management/archiving. Authenticity. Digital forensics. Strategies and tools for assessing, deduplicating, and accessioning e-records. PeDALS. Duke Data Accessioner. Archivematica. Institutional repositories for records archival.*

"Digital records preservation: where to start guide." <http://isotc.iso.org/livelink/livelink?func=ll&objId=10083866&objAction=Open&nexturl=%2Flivelink%2Flivelink%3Ffunc%3D11%26objId%3D8800147%26objAction%3Dbrowse%26sort%3Dname>

Briston, Heather, and Karen Estlund. "From passive to active preservation of electronic records." *Ariadne* 65 (2010). <http://www.ariadne.ac.uk/issue65/briston-estlund/>

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. "Digital forensics and born-digital content in cultural heritage collections." <http://www.clir.org/pubs/reports/pub149/pub149.pdf>

PeDALS. "About PeDALS." <http://www.pedalspreservation.org/About/Default.aspx> (please read through all the sub-pages under the About menu)

"Data accessioner." <http://library.duke.edu/uarchives/about/tools/data-accessioner.html>

"Archivematica." <http://archivematica.org/> and <http://archivematica.org/wiki/index.php?title=Overview>

### **Week of May 2: Existing data archives. Data discovery.**

*Learning objectives: Finding disciplinary data archives, open and subscription. Licensing issues with data archives. Data, the ILS, and discovery layers. Google and data. Data reuse. Hathi Trust. Digital libraries as humanities-data archives. Journals and supplementary data.*

Berman, Francine. "We need a research data census." *Communications of the ACM* 53:12 (2010). <http://dx.doi.org/10.1145/1859204.1859220>

Carleton College. "Data, Datasets, and Statistical Resources." <http://gouldguides.carleton.edu/content.php?pid=65030&sid=480389> (please look through all the tabs)

# Assignments

## FINAL PROJECT

For your final project, you will work in a group to help solve a data-curation problem for a campus entity (faculty member, department, or research unit). You will determine the nature and extent of the problem, make a plan to solve it, agree with your client and me about how much of the problem your group can solve over the course of the semester, and work to the resulting plan.

Due dates:

- Project plan: February 14. This should be a plan for solving the *entire problem* as presented to you by the client. Don't worry; you will not necessarily be expected to complete the entire plan in one short semester! To construct this plan, you should approach the client to perform a data interview. A data-curation profile should form part of this plan as well. You may decide to revise this plan over the course of the semester!
- Project checkins: March 7 and April 18. On the project-discussion forum in Learn@UW, please let me know what you're doing and how it's going.
- Project wrapup: May 9. A BRIEF (six pages is too many; two might be enough) statement of the problem presented, the nature of the solutions suggested and deployed, the progress made over the semester, and any larger issues brought to light during the process. If you have revised your project plan, please provide a copy of the revised plan as well (this does not count against your pagecount).

I will check with clients about your group's professionalism, competence, and accomplishments before I assign grades.

Grading rubric:

- Project plan: 40% (understanding of the problem, appropriateness of suggested solutions, clear expression)
- Checkins: 20% (evidence of good project management, clear expression)
- Wrapup: 40% (accomplishment, overcoming obstacles, professional relationship with client, clear expression)

**On group projects:** The idea that group projects are uniquely designed to torture library school students is a snare and a delusion. Librarianship (as well as data curation specifically) includes immense amounts of collaborative work, from local committees and task forces to involvement in national professional organizations and everything in between. None of the obstacles to working in groups – scheduling, free riders, personality conflicts – disappears when you receive your degree. If you are not good at working in a team, now is the time to learn!

## Well-formed XML instance

You will find a non-well-formed instance of XML on Learn@UW. Make it well-formed. Those of you with existing XML experience may wish to know that this instance purported to be TEI (<http://tei-c.org/>); by all means make it a valid (if bare-bones) TEI instance.

Grading rubric: will it parse? Due January 24.

## Current awareness via blogs

Learn@UW will have a list of relevant blogs. Choose one and browse its archive. When you find a substantive post that interests you, post a link and summary, plus any reactions or questions you have, to the appropriate forum on Learn@UW.

Grading rubric: did you find a good post? did you understand it? did you say anything useful about it? Due February 21.

## NSF data-management plan

A real-world example (anonymized!) of an NSF data-management plan will be posted to Learn@UW. Read it, find the correct NSF guidance for it, evaluate the plan according to that guidance and your own sense of what is necessary, and post your evaluation to Learn@UW. Please write your evaluation as though the NSF grant applicant were going to read it. Due February 28.

Grading rubric: did you make appropriate recommendations? was your expression professional?

## Outlining a data-curation program

Imagine that you are charged with improving the state of research-data curation on the UW-Madison campus. Ask questions in the weekly Learn@UW forum that will allow you to suggest organizational structures and funding sources for such a program. You are NOT bound by what this campus is currently doing! Due March 28.

Grading rubric: what did you add to the discussion?

### **Crawling a website for archival**

You will be given a website URL on Learn@UW. Using the software of your choice, capture the entirety of this website (n.b. not just its home page!) for archival. Make sure you proof your capture against the original and fix any problems! Zip your results and put them in the designated Learn@UW dropbox. On the weekly Learn@UW forum, reflect on any problems you encounter, as well as the fitness of the website for long-term preservation. What might you do to improve that fitness?

Grading rubric: did you miss anything? did you make useful observations about the site?

### **METS file**

You will choose a digital object from a list available on Learn@UW. Over the semester (various due dates), you will construct a valid METS file for it, including descriptive, structural, and technical metadata. Put the final file in the designated Learn@UW dropbox when it is completely finished (April 18).

Grading rubric: did you describe the object accurately? is the file valid METS?

### **Evaluating software platforms**

Various sorts of software will be made available on the class server, as well as pointers to demo installations elsewhere. Using the same digital object you chose for your METS assignment, evaluate TWO software packages (from different software categories, please) as to:

- Whether they can ingest the object at all.
- Whether they can usefully ingest and work with your METS metadata file.
- Whether they can present the object to an end-user appropriately and usefully.
- How much configuration or customization would be needed for the software to be useful for the object.

Post your answers to the Learn@UW forum, along with explanations of where the software package fell short (if it did). You are welcome to answer some or all of these questions by finding similar objects in real-world digital libraries, repositories, or archives. Of course you can also use software documentation.

Grading rubric: are your answers correct? have you thought through the possibilities and challenges?