Frameworks for a Data Management Curriculum

Course plans for data management instruction to undergraduate and graduate students in science, health sciences, and engineering programs.

Developed by the Lamar Soutter Library, University of Massachusetts Medical School and the George C. Gordon Library, Worcester Polytechnic Institute

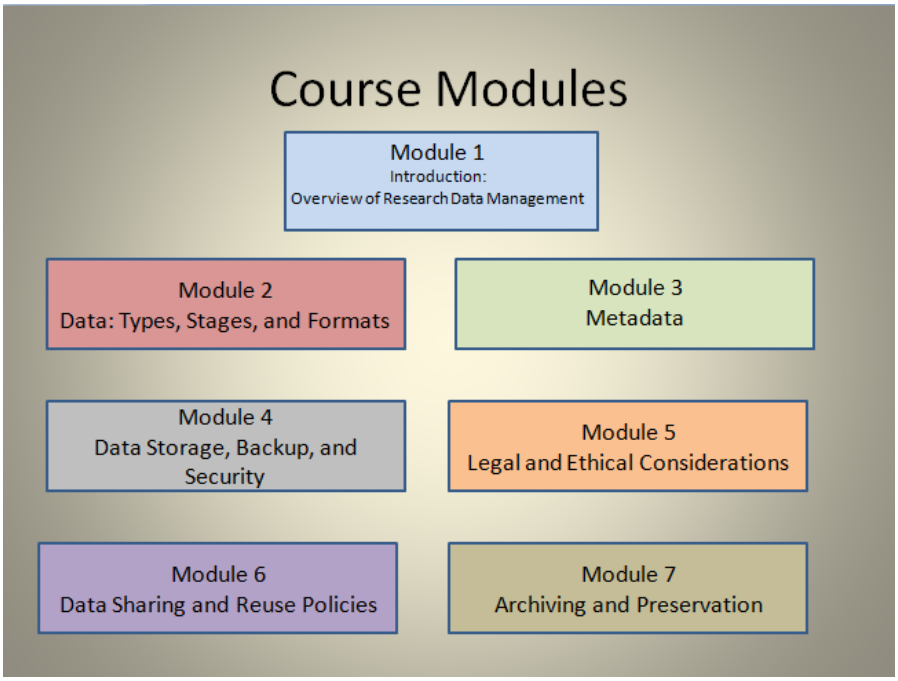
This project is made possible by a grant from the U.S. Institute of Museum and Library Services and with funds from the National Library of Medicine under Contract No. N01-LM-6-3508.

Co-PIs: Elaine Martin, DA and Tracey Leger-Hornby, Ph.D.
Project Coordinator: Donna Kafel, MLIS

February 2012

Introduction

The *Frameworks for a Data Management Curriculum* packet has been developed for teaching research data management to undergraduate and graduate level students in the sciences, health sciences, and engineering disciplines. The curriculum has been designed as a series of seven course modules in order to allow maximum flexibility for customizing instruction. With this framework model, faculty have the option to integrate the entire series of modules into a program of study or select individual modules that target their students' learning needs.



Included within the frameworks are:

- The curriculum's connection to National Science Foundation data management plan requirements
- Lesson plans for Modules 1-7
- Summary list of readings for the modules
- A simplified data plan for student exercises

- Summary of teaching points from Research Data Management Cases
- 4 Research Data Management Cases
- Course content for Module # 5 (pilot module)
- Excerpts of the 4 research case studies for Module #5 (pilot module)
- Assessments

The frameworks include four research case studies that illustrate data management concepts in various science and medical research settings including clinical medical research (Case A: Outcomes from Orthopedic Implant Surgery), biomedical lab research (Case B: Regeneration of Functional Heart Tissue), qualitative behavioral health research (Case C: Improving End-of-Life Care for African Americans) and aerospace engineering research (Case D: Characterizing a Component of a Rocket Engine Used to Control Satellites in Orbit). Each research case is preceded by a summary of its teaching points. At the end of each case study are overview discussion question(s), and suggested discussion questions relevant to each of the seven modules. The research case studies can be used for class or small group discussion and as an assessment tool for a module.

Content for Course Module #5 on Legal and Ethical Considerations for Research Data has been fully developed as a proof of concept. Along with the class content are excerpts of the four research case studies A-D that illustrate legal and ethical issues in research data management.

The final section of the frameworks (pp. 59-66) includes assessment questions and answers for the excerpts of the research cases for module #5.

These materials are presented for use by faculty and librarians. Contact Donna Kafel, Project Coordinator, at Donna.Kafel@umassmed.edu with questions, feedback, or to request further information about the frameworks and how they were developed.

Data Management Curriculum Frameworks
Table of Contents

Connection to the National Science Foundation Data Management Plan Requirements

Each of the seven course modules in this data management curriculum address one or more of the following components of the National Science Foundation's requirement for Data Management Plans for NSF funded research projects from the National Science Foundation Proposal and Award Policies Procedures Guide: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp and the National Science Foundation Directorate of Biological Sciences http://www.nsf.gov/bio/pubs/BIODMP061511.pdf

1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;

2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);

3. physical and/or cyber resources and facilities (including third party resources) [that]will be used to store and preserve the data;

4. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;

5. policies and provisions for re-use, re-distribution, and the production of derivatives;

6. plans for archiving data, samples, and other research products, and for preservation of access to them.

Lesson Plan for Module # 1- Overview of Research Data Management	
Learning Objectives	By participating fully in this class, student will be able to: 1. Explain what research data is 2. Explain the need for managing/sharing research data and identify relevant public policies 3. Explain the lifecycle continuum to manage and preserve research data 4. Understand that data should be managed differently in different phases of the life cycle 5. Be familiar with data management plan (DMP) requirements used to characterize and plan for the lifecycle of research data. 6. Identify the value and relative importance of data management to the success of a research project.
Lecture Content	1. Explain broadly what is research data 2. Illustrate need for proper data management practices. Present 3 examples (research data from student project, federally funded project at local institution, and from multi-institutional projects of national scope). 3. Describe funding agency requirements for data management 4. Describe research data lifecycle continuum phases: create, process, analyze, preserve, give access, reuse. Show diagram, and use real life example to illustrate each phase of the continuum. http://www.data-archive.ac.uk/create-manage/life-cycle 5. Show and compare sample data management plan requirements, give examples of DMPs for different funding agencies http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/resources.jsp#a02 6. Present simplified data management plan template.
Activities	1. Identify data sets collected and/or generated in examples used in lecture content #2 above 2. Create a data management plan for one of the cases using the simplified data management plan template on page 21.
Assessment	Read excerpt from research data management case #A: Outcomes from Orthopedic Implant Surgery (Illustrates the challenges in conducting a multiyear research project with changing personnel each year) and respond to questions.
Readings	1. Promoting the Stewardship of Research Data, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (2009): pages 95-99 http://books.nap.edu/openbook.php?record_id=12615&page=95 2. Why Share Data? UK Data Archive http://www.data-archive.ac.uk/create-manage/planning-for-sharing/why-share-data 3. Introduction: A Revolution in Science: p.11-13 from Harnessing the Power of Digital Data for Science and Society (2009) http://www.nitrd.gov/about/harnessing_power_web.pdf 4. Steps in the Research Life Cycle, Scientific Data Consulting, University of Virginia Library http://www2.lib.virginia.edu/brown/data/lifecycle.html 5. Data Management and Publishing http://libraries.mit.edu/guides/subjects/data-management/funding.html 6. Funding Agency and Data Management Guidelines: http://www.lib.umn.edu/datamanagement/funding 7. Example Data Management Plan http://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf .

Lesson Plan for Module #2- Types, Formats, and Stages of Data	
Learning Objectives	By participating fully in this class, student will be able to: 1. Explain what a research data set is and the range of data types 2. Identify stages of research data 3. Identify common potential storage formats for data that will be accessible in the future and non-proprietary where possible (i.e., not related to proprietary or custom software/instruments used for capturing/analyzing data) 4. Identify relevant quality control techniques/technical standards 5. Identify methods of recording data that are specific to student's discipline and research interests. 6. Define data collection recording policies/procedures for student's research.
Lecture Content	1. Explain what research data is a. Discuss the various types of research data: quantitative (experimental measures), survey results, observations, data generated by simulation/test models, qualitative (text, images, video), specimens, existing data (source). b. Associate common types of data with major disciplines. 2. Review lifecycle continuum phases of research data and identify data stages as they relate to the continuum: raw data, processed data, analyzed, finalized and/or published data. Refer to UK Data Archive model of research data lifecycle http://www.data-archive.ac.uk/create-manage/life-cycle 3. Discuss storage file formats (e.g. Excel, Access, STATA, SAS etc.) and the pros and cons of each for the short and long term. 4. Discuss and review the importance of documenting format migration history (microscope to Excel to PowerPoint). 5. Discuss and review the importance of consistent data collection recording practices to maintain quality and standardization of data. Review examples by discipline, such as data dictionary, lab notebook.
Activities	1. Using sample dataset or case study, identify potential file formats. 2. Match data examples to appropriate lifecycle continuum phases and data stages. 3. Two sets of data one organized and well documented; the other not. Students analyze data and discuss the process and issues encountered.
Assessment	Quiz on lecture content
Readings	1. Defining Research Data (University of Edinburgh) http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-mgmt/research-data-definition 2. File Formats for Long-Term Access (MIT data management guide): http://libraries.mit.edu/guides/subjects/data-management/formats.html 3. Create and Manage Data: Formatting your Data http://www.data-archive.ac.uk/create-manage/format

Lesson Plan for Module #3- Contextual Details Needed to Make Data Meaningful to Others	
Learning Objectives	By participating fully in this class, student will be able to: 1. Understand what metadata is 2. Understand why metadata is important 3. Identify applicable standards for documenting and capturing metadata 4. Understand disciplinary practices associated with the collection and sharing of metadata 5. Identify an approach to creating metadata for a project
Lecture Content	1. Explain what metadata is (use examples) 2. Explain how metadata facilitates definition of data structure, ownership, reuse, accessibility, discoverability 3. Review the importance of metadata standards in relation to collaboration and sharing. Identify sources of metadata standards (provide examples) 4. Discuss the application of national standards in the local setting by using clear and consistent descriptions (for example, controlled vocabulary, date format, etc.) and naming conventions so data may be located and used by others 5. Present and review a variety of metadata elements to allow your research data sets to be managed, preserved, and accessible to others. Such as: a. Basic project info: project name, funders, budget, duration, partner organizations, data creator, creator's institution, discipline/sub-discipline of research area focus, data owners, purpose of research, staffing and roles b. For each data set, data type, stage of data, stage(s) most valuable for preservation (ingest) c. Data collection methods d. Recording creation date/time for data files e. Instruments and/or software used to create/process data (including instrument or software version) f. Quality assurance, validation strategies used g. Data volume, file formats, number of files h. Data organization – directory structure, file naming conventions, file structure i. Content of data – variable names and descriptions, classification schemes used j. Data analysis methods; algorithms used to process data k. Confidentiality, access and use conditions 6. Discuss approaches to creating metadata for a research project a. How/who will create metadata? b. Possibilities for automating metadata creation c. Which metadata standard will be used or will a local schema be created? How will the metadata be associated with the research data?
Activities	1. Using a research case, identify the basic project information 2. Transfer basic project information to metadata template (template will be provided) 3. Review sample repository record (web display view and the metadata files that make up the web view). (IR record will be provided)
Assessment	Quiz on lecture content

Lesson Plan for Module #3- Contextual Details Needed to Make Data Meaningful to Others: Readings

Readings	1. File Naming Conventions from the University of Minnesota http://researchdata.wisc.edu/manage-your-data/file-naming-and-versioning/ 2. Version Control and Authenticity http://www.data-archive.ac.uk/create-manage/format/versions 3. Video: What is Metadata (less than 5 minutes) http://vimeo.com/3161893 4. Introduction to Metadata: Setting the Stage (Getty Research Institute) http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html 5. Documentation and Metadata (MIT Libraries) http://libraries.mit.edu/guides/subjects/data-management/metadata.html 6. Seeing Standards: A Visualization of the Metadata Universe http://www.dlib.indiana.edu/~jenlrile/metadatamap/

Lesson Plan for Module #4- Data Storage, Backup, and Security

Learning Objectives	By participating fully in this class, student will be able to: 1. Understand why data storage, backup, and security of research data are important 2. Understand data storage, backup, and security methods for research data 3. Understand best practices for research data storage, access control, migration to newer storage media, and security of research data 4. Identify an approach to creating a data storage, backup, and security plan for a project
Lecture	1. Place data storage, backup, and security in the context of data management plans 2. Define data storage and review current practices and the pros and cons of each, e.g.: a. Describe the types of media used for storage and its storage capacity, longevity, retrieval effectiveness and ease of upgrade to newer media b. Explain the potential need to migrate data files to new platforms and standards c. Review the importance of assigning responsibility for storing and backing up data 3. Explain data backup strategies and review current practices and the pros and cons of each, e.g.: a. Explain the importance of estimating the length of time your data needs to be stored or preserved b. Discuss need to preserve full naming conventions for backup files c. Communicate options for employing a timely back-up process for data to a media that has a high level of stability and interoperability 4. Review different levels of security (access, data integrity, system protection) and related issues a. Identify ways of protecting access to your data: 1) Unique User ID/Password 2) Provide access through a centralized system 3) Limiting of administrator access rights 4) Limitations of wireless devices to protect access b. Explain multiple ways of protecting your computer systems: 1) Updated anti-virus software. 2) Up-to-date versions of client software and media storage devices. 3) Use of a firewall. 4) Use of intrusion detection software to monitor access. 5) Restrict physical access to computers and media c. Describe methods for protecting data integrity: 1) Use of encryption, electronic signatures, watermarking for authorship verification and changes made to data files. 2) Regular back-up schedule for data files offsite (use of secondary storage sites) d. Discuss strategies for destruction of data (especially confidential data) if needed
Activities	1. How could the backup process be improved for research data management case B? 2. How could the security of lab notebooks be improved for research data management case B?
Assessment	Read excerpt of research data management case C: Improving End-of-Life Care for African Americans, and respond to questions at end.

Lesson Plan for Module #4- Data Storage, Backup, and Security: Readings

Readings	1. Backing Up Data from the UK Data Archive: http://www.data-archive.ac.uk/create-manage/storage/back-up 2. The University of Edinburgh's Guide to Data Sharing and Preservation: http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-sharing/preservation 3. Information Security Primer: http://www.seattle.gov/informationSecurity/pdf/EPRI_securityPrimer.pdf (Security overview Section 3, Section 6.1.2: Identification and Authentication of Users, Section 6.1.3: Cryptography) 4. NASA networks open to cyber attacks: http://www.net-security.org/secworld.php?id=10824

Lesson Plan for Module #5- Legal and Ethical Considerations for Research Data	
Learning Objectives	By participating fully in this class, student will be able to: 1. Explain ownership considerations related to data sharing 2. Explain and evaluate potential legal issues connected to your data; intellectual property, copyright claims, licenses needed for use, monetary charges for data 3. Explain ethical considerations related to data sharing 4. Understand privacy levels for research data as required by potential funding agencies 5. Recognize the importance of privacy with some forms of research data (HIPAA) 6. Understand the importance of removing key personal identifiers to facilitate confidentiality 7. Understand the need for data attribution and citation.
Lecture Content	1. Who owns research data? 2. What ethical considerations does one need to be aware of when using another's data or sharing data? o What is the importance of acknowledging the source of data that are used in research? o What are the components of a data citation? 3. How do intellectual property laws and/or copyright protections relate to research data and researchers? 4. Issues related to licensing or charging for reuse of research data 5. Privacy Considerations o Privacy requirements of funding agencies related to reuse o How do HIPAA requirements relate to reuse of research data? o How can informed consent either protect against or allow for reuse of data? o De-identification of data for re-use
Activities	1. Read scenario and discuss answers: Who owns the data? (lab scenario of graduate student wanting to take data) (from Columbia Responsible Conduct of Research) http://ori.dhhs.gov/education/products/columbia_wbt/rcr_data/case/index.html#2 2. Locate and review Intellectual Property Policy of your local Institution. 3. Locate and review the website of your local Institutional Review Board to see what information it has posted regarding data management (e.g. patient privacy, support for writing data management plans). 4. Identify components of a data citation.
Assessment	Have students read excerpt of research data management case A (Outcomes from Orthopedic Implant Surgery) or research data management case C (Improving End-of-Life Care for African Americans) 1. Read excerpt of scenario and respond to questions 2. Read and discuss commentary "Henrietta's Dance" http://www.jhu.edu/jhumag/0400web/01.html or view "Henrietta Lack-CBS Sunday Morning" http://www.youtube.com/watch?v=wRrNjHYxP_o&feature=related

Lesson Plan for Module #5- Legal and Ethical Considerations for Research Data:
Readings

Readings	1. Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule http://privacyruleandresearch.nih.gov/pr_02.asp 2. Guidelines for Responsible Data Management in Scientific Research http://ori.hhs.gov/education/products/clinicaltools/data.pdf pgs. 6-8 3. "Who Owns Research Data?" http://ori.dhhs.gov/education/products/columbia_wbt/rcr_data/case/index.html#2 4. "Constructing Access Permissions", University of Oregon Libraries: http://libweb.uoregon.edu/datamanagement/sharingdata.html#three 5. Prison for HIPAA Privacy Violator Health Data Management Magazine, 06/01/2010 http://www.healthdatamanagement.com/issues/18_6/hipaa-prison-for-hipaa-privacy-violator-40382-1.html 6. International Polar Year Data and Information Service: How to Cite a Data Set http://ipydis.org/data/citations.html 7. Ball, A. & Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides 8. Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4), http://www.dlib.org/dlib/march07/altman/03altman.html 9. Green, T. (2009). We need publishing standards for datasets and data tables. *OECD Publishing White Papers*, OECD Publishing, http://www.oecd.org/document/25/0,3746,en_21571361_33915056_42600857_1_1_1_1,00.html

Lesson Plan for Module #6- Data Sharing & Re-Use Policies	
Learning Objectives	By participating fully in this class, student will be able to: 1. Identify issues/obstacles related to re-use and sharing 2. Understand publisher's and licensing restrictions on re-use of data and analysis software and instrumentation 3. Understand Open Access requirements 4. Understand controversies surrounding open science, open data 5. Address re-use/sharing requirements from granting agencies or sponsors 6. Address the need for conversion to standard formats needed for re-use 7. Understand different types of collaborative workspaces for sharing data 8. Identify who can share/access your data and for what purpose 9. Determine requirements for pre/post publication access for project phases of the research 10. Determine temporary or permanent access policy 11. Define process steps and access levels for gaining access 12. Understand options for maximizing data reuse
Lecture Content	1. Overview of issues/obstacles related to re-use and sharing 2. Review potential legal restrictions on re-use (e.g., copyright, IP, patents, proprietary/commercial) (Refer to course module 4) 3. Publishers restrictions vs. Open Access policies 4. Controversies surrounding open science and open data 5. Review requirements & restrictions from funding agencies 6. Review of current standard formats for data re-use (Refer to course module 1) 7. Virtual Research Environments 8. Staging repositories for research data 9. Pros and cons of sharing preliminary data (non-published) be shared with collaborators and/or the public on an immediate basis (researcher vs. tax payer perspective) 10. Present examples of temporary and permanent access policies (e.g., patent challenge) 11. Institutional, government, licensing determinants for gaining access - https://cnda.wustl.edu/ 12. Examples of process steps and access levels for collaborator and/or public use 13. Use of Creative Common License and Science Commons Database protocol to maximize data re-use - Creative Commons website
Activities	1. Investigate open access policy at your institution 2. Compare policies of restrictive publisher to non-restrictive publisher (non-open access) 3. Identify potential re-users of data in your research area, the value of your research data for re-use, and a dissemination strategy 4. Discuss obstacles you might encounter to sharing your research data 5. Develop a statement to maximize re-use of your data
Assessment	Research data management case D Aerospace engineering case: "Characterizing a Component of a Rocket Engine used to Control Satellites in Orbit" and related quiz questions

Lesson Plan for Module 6: Data Sharing and Re-use Policies: Readings

Readings	1. Why data-sharing policies matter http://www.pnas.org/content/106/40/16894.full 2. Alan E. Guttmachera, Elizabeth G. Nabel and Francis S. Collin Data Sharing and Consent March 2010 By Ciara Curtin http://www.genomeweb.com/data-sharing-and-consent 3. Data Ownership from Responsible Conduct in Data Management Faculty Development and Instructional Design Center - Northern Illinois University http://ori.dhhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html 4. Data-Sharing Dilemmas: Allowing Pharmaceutical Company Access to Research Data JR Anderson… - IRB: Ethics & Human Research, 2009 - thehastingscenter.org ISCB Public Policy Statement on Open Access to Scientific and Technical Research Literature http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002014 5. Overview of Scientific Data Sharing and Reuse Policies of the Federal Government at The Value of Shared Access and Reuse of Publicly Funded Scientific Data, 2010 http://sites.nationalacademies.org/PGA/brdi/PGA_059258 6. Open Data and the Social Contract of Scientific Publishing . OECD Publishing White Papers, OECD Publishing, http://www.jstor.org/stable/10.1525/bio.2010.60.5.2 .

Lesson Plan for Module #7- Plan for Archiving and Preservation of Data	
Learning Objectives	By participating fully in this class, student will be able to: 1. Explain options for a long-term sustainable preservation strategy/policy for your data (eg, discipline specific, institutional, departmental). 2. Identify types of repositories/archives (discipline-based, institutional, etc.) 3. Choose appropriate subject repository for long term storage of data 4. Understand process issues for depositing data in repository 5. Identify issues related to discovery of relevant data sets for re-use 6. Understanding the need for querying and retrieval methods - discovery aids for multiple user communities to find the data they want to re-use 7. Explain data management tools and services available 8. Understand costs for data storage, management tools and services
Lecture Content	1. Present examples of different options for preservation strategies/policies 2. Types of repositories available and factors influencing where to deposit data (present examples of discipline specific and institutional repositories) 3. Review the costs and benefits associated with using public data repositories 4. Present examples of submission guidelines (http://conservancy.umn.edu/UDCsubmissionguidelines.pdf) 5. Issues related to searching for relevant data sets across repositories and disciplines 6. Retrieval issues & methods – getting data from the repository 7. Types of data management tools and services (example: Discovery Garden Services - http://www.discoverygarden.ca/node/13) 8. Business model for long-term archiving of data (DataSpace at Princeton)
Activities	1. Review institutional policies for data archiving and preservation 2. Review submission processes for a discipline-specific repository 3. Review paper and discuss model for charging back data hosting (http://arks.princeton.edu/ark:/88435/dsp01w6634361k)
Assessment	Quiz on lecture content
Readings	1. University of Oregon's Managing your Data: Data Centers and Repositories http://libweb.uoregon.edu/datamanagement/repositories.html 2. Business Model and Cost Estimation: DRYAD Repository Case Study http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf 3. DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data http://net.educause.edu/LIVE1024 4. The DataVerse Network http://dlib.org/dlib/january11/crosas/01crosas.html

Cumulative List of Readings for Data Management Curriculum Course Modules

Module 1: Overview of Research Data Management

1. Promoting the Stewardship of Research Data, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (2009): pages 95-99 http://books.nap.edu/openbook.php?record_id=12615&page=95
2. Why Share Data? UK Data Archive http://www.data-archive.ac.uk/create-manage/planning-for-sharing/why-share-data
3. Introduction: A Revolution in Science: p.11-13 from Harnessing the Power of Digital Data for Science and Society (2009) http://www.nitrd.gov/about/harnessing_power_web.pdf
4. Steps in the Research Life Cycle, Scientific Data Consulting, University of Virginia Library http://www2.lib.virginia.edu/brown/data/lifecycle.html
5. Data Management and Publishing http://libraries.mit.edu/guides/subjects/data-management/funding.html
6. Funding Agency and Data Management Guidelines: http://www.lib.umn.edu/datamanagement/funding
7. Example Data Management Plan http://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf

Module 2: Types, Formats, and Stages of Data

1. Defining Research Data (University of Edinburgh) http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-mgmt/research-data-definition
2. File Formats for Long-Term Access (MIT data management guide): http://libraries.mit.edu/guides/subjects/data-management/formats.html
3. Create and Manage Data: Formatting your Data http://www.data-archive.ac.uk/create-manage/format

Module 3: Contextual Details Needed to Make Data Meaningful to Others

1. File Naming Conventions from the University of Minnesota http://researchdata.wisc.edu/manage-your-data/file-naming-and-versioning/
2. Version Control and Authenticity http://www.data-archive.ac.uk/create-manage/format/versions
3. What is Metadata? (less than 5 minutes) http://vimeo.com/3161893
4. Introduction to Metadata: Setting the Stage (Getty Research Institute) http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html
5. Documentation and Metadata (MIT Libraries) http://libraries.mit.edu/guides/subjects/data-management/metadata.html
6. Seeing Standards: A Visualization of the Metadata Universe http://www.dlib.indiana.edu/~jenlrile/metadatamap/

Module 4: Data Storage, Backup, and Security

1. Backing Up Data from the UK Data Archive: http://www.data-archive.ac.uk/create-manage/storage/back-up
2. The University of Edinburgh's Guide to Data Sharing and Preservation: http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-sharing/preservation
3. Information Security Primer: http://www.seattle.gov/informationSecurity/pdf/EPRI_securityPrimer.pdf (Security overview Section 3, Section 6.1.2: Identification and Authentication of Users, Section 6.1.3: Cryptography)
4. NASA networks open to cyber attacks: http://www.net-security.org/secworld.php?id=10824

Module 5: Legal and Ethical Considerations for Research Data

1. Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule http://privacyruleandresearch.nih.gov/pr_02.asp
2. Guidelines for Responsible Data Management in Scientific Research http://ori.hhs.gov/education/products/clinicaltools/data.pdf pgs. 6-8
3. "Who Owns Research Data?" http://ori.dhhs.gov/education/products/columbia_wbt/rcr_data/case/index.html#2
4. "Constructing Access Permissions", University of Oregon Libraries: http://libweb.uoregon.edu/datamanagement/sharingdata.html#three
5. Prison for HIPAA Privacy Violator Health Data Management Magazine, 06/01/2010 http://www.healthdatamanagement.com/issues/18_6/hipaa-prison-for-hipaa-privacy-violator-40382-1.html
6. International Polar Year Data and Information Service: How to Cite a Data Set http://ipydis.org/data/citations.html
7. Ball, A. & Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides
8. Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4), http://www.dlib.org/dlib/march07/altman/03altman.html
9. Green, T. (2009). We need publishing standards for datasets and data tables. *OECD Publishing White Papers*, OECD Publishing, http://www.oecd.org/document/25/0,3746,en_21571361_33915056_42600857_1_1_1_1,00.html

Module 6: Data Sharing & Re-Use Policies

1. Why data-sharing policies matter http://www.pnas.org/content/106/40/16894.full
2. Alan E. Guttmachera, Elizabeth G. Nabel and Francis S. Collin Data Sharing and Consent March 2010 By Ciara Curtin http://www.genomeweb.com/data-sharing-and-consent

3. Data Ownership from Responsible Conduct in Data Management Faculty Development and Instructional Design Center - Northern Illinois University http://ori.dhhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html
4. Data-Sharing Dilemmas: Allowing Pharmaceutical Company Access to Research Data JR Anderson… - IRB: Ethics & Human Research, 2009 - thehastingscenter.org
   ISCB Public Policy Statement on Open Access to Scientific and Technical Research Literature
   http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002014
5. Overview of Scientific Data Sharing and Reuse Policies of the Federal Government at The Value of Shared Access and Reuse of Publicly Funded Scientific Data, 2010
   http://sites.nationalacademies.org/PGA/brdi/PGA_059258
6. 6. Open Data and the Social Contract of Scientific Publishing
   http://www.bioone.org/doi/pdf/10.1525/bio.2010.60.5.2

Module 7: Plan for Archiving and Preservation of Data

1. University of Oregon's Managing your Data: Data Centers and Repositories
   http://libweb.uoregon.edu/datamanagement/repositories.html
2. Business Model and Cost Estimation: DRYAD Repository Case Study
   http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf
3. DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data http://net.educause.edu/LIVE1024
4. The DataVerse Network
   http://dlib.org/dlib/january11/crosas/01crosas.html

Simplified Data Management Plan

Based on the NSF requirements for a data management plan, this simplified data management plan can be used as a template for creating a data management plan for Module 1 Activity.

1. Types of data
    a. What types of data will you be creating or capturing? (experimental measures, observational or qualitative, model simulation, existing)
    b. How will you capture, create, and/or process the data? (Identify instruments, software, imaging, etc. used)

2. Contextual Details (Metadata) Needed to Make Data Meaningful to others
    a. What file formats and naming conventions will you be using?

3. Storage, Backup and Security
    a. Where and on what media will you store the data?
    b. What is your backup plan for the data?
    c. How will you manage data security?

4. Provisions for Protection/Privacy
    a. How are you addressing any ethical or privacy issues (IRB, anonymization of data)?
    b. Who will own any copyright or intellectual property rights to the data?

5. Policies for re-use
    a. What restrictions need to be placed on re-use of your data?

6. Policies for access and sharing
    a. What is the process for gaining access to your data?

7. Plan for archiving and preservation of access
    a. What is your long-term plan for preservation and maintenance of the data?

Case A: Outcomes from Orthopedic Implant Surgery

Summary of Teaching Points

Module 1: Overview of Research Data Management

- the challenges in conducting a multiyear research project with changing personnel each year

Module 2: Types, Formats & Stages of Data

- Proprietary software using file naming conventions that were not clear and not under control of investigator and software stored data in its own application database that needed to be exported to a common format for storage
- The need to update proprietary software as new releases become available to maintain support from vendor and keep data formats current

Module 3: Contextual Details

- No validity checks on data entry for patient survey data

Module 4: Data Storage, Backup and Security

- No plan for storage of survey source documents
- Security and backup plan in place

Module 5: Legal and Ethical Issues

- Use of novel instrumentation with single license proprietary software and need to use it on multiple PCs
- Need to de-identify patient data when doing research with human subjects
- Need for IRB application and informed consent form

Module 6: Data Sharing and Re-Use

- Informed consent restrictions

Module 7: Plan for Archiving and Preservation of Data

- None

Research Data Management Case A: Outcomes from Orthopedic Implant Surgery

Dr. X wrote a 5 page proposal for funding for a study to use a novel monitor with proprietary software to assess patient outcomes 2 years after orthopedic implant surgery. This prospective longitudinal study would determine the rate of sub-optimal outcomes based on specialized analysis using the proprietary software that accompanied the monitor. The study was funded and the research resident working with the PI prepared the IRB application that received approval. With a clearly defined research hypothesis, innovative monitor technology, and IRB application and consent form complete, the goal was to collect the same measures over 3 years. The resident on the project began to enroll patients and collect data on his office PC. At the end of the training year, the resident handed the study to the next resident whose responsibility was to continue enrollment and collect the one year follow-up data on the initial cohort. At the end of study year 2, the third resident continued to enroll more patients, collect 2 year outcomes from the first cohort and 1 year outcomes on the second cohort. A very large volume of data had been collected and the new research resident was responsible for integrating and analyzing the data in preparation for publication the following year. She encountered a series of data issues that were not documented or clear to her. While the PI had originally defined the data to collect, she had not been directly involved in the data collection and could not answer the questions. The first resident who started the project had completed training and left the institution.

The same patients were followed for three years so it required tracking them down to have them come in to allow for collection of data via multiple sources: patient surveys, accelerometer measurements, and surgeon notes from the physical exams. The Principal Investigator had HIPAA authorization to use the patient's name, Med record #, and telephone/address to contact them for follow-up. However, the data base was organized by unique study ID assigned to each patient.

The study was complex due to the need to collect and integrate data from these three different sources:

1) Patient-generated data regarding their demographics and their symptoms, the amount of pain and disability. Patients filled out a hand-developed paper survey at baseline and annually for 3 years. The core outcome measures were the same from the survey each year, but the basic demographics questions were not repeated each year. Much of it was based on pre-existing standardized forms so there were already some data definitions for some responses. The format was a mix of these standardized questions (well tested and validated responses) and some new questions with uncertain responses (open-ended response option). Data were hand-entered into an Excel spread sheet; so there was no application of data quality checks (number range, etc) as data were entered. Survey data were entered by various people into an excel spreadsheet and the source documents were stored in multiple locations. Eventually the patient surveys were moved onto a direct computer data entry system to avoid the validation problem. The data were captured in survey software that could be downloaded into a spreadsheet/data file for analysis.

2) The second source was measurements from an accelerometer that did 24 hour tracing of patients' steps and walking rate annually. This novel monitor came with proprietary software that produced bulk summary statistics on an excel spreadsheet. However, the study required individual patient records that had to be exported for analysis. We exported the data on each patient from the software to a data file. The monitor analytic software was on a lab PC originally. It was a proprietary software package that could be loaded only on one computer and it had to be handed off as residents changed. The rest of the data from other sources were on the research assistant's PC. We bought another monitor software license to get it off the original PC because the monitor analytic data were housed there and we then put it on a laptop. The specialized monitor analysis software used naming conventions that were not clear and data were stored in the proprietary software. The software itself was updated across the 3 years.

3) The third source was a surgeon note in the EMR and there was no standard for this surgeon note resulting in varied styles of documentation. Residents read the charts every month related to patients in the study to identify any follow-up MD office visits and to extract physical exam measures which were inserted into a structured database with data definitions for each measure.

The data from these multiple sources needed to be integrated for a biostatistician to apply longitudinal modeling software. ACCES was used as the final data base and was used to house the total data set and integrate data (through a flat file) from all the sources. Data sub-sets were imported to STATA software for particular analyses, as needed. Data were stored on a server solely for research that was password protected, backed up nightly, and protected by institutional firewalls, etc. (not on a computer). STATA software was used for data analysis such as linear and logistic multi-variate models. Backup was done nightly through the institutional IS procedures for data stored on their research servers. Security measures such as passwords, limited access, firewall, etc. were used to safeguard the data.

Module 1 (Overview module) discussion question:

What issues need to be addressed on this project related to the 7 segments of the data management plan components?

Discussion Questions for Other Modules:

1. Types of data
    a. What types of data are being collected for this study?
    b. How will you ensure all research assistants/residents used the same data sources and data definitions?
    c. What would be needed in a data management plan to describe use of novel equipment?
    d. What needs to be in the plan related to the patient survey data capture and the capture of surgeon notes?
    e. What analytical methods and mechanisms will be applied to your data either prior to or post integration
    f. What type of outcome data will be generated?

2. Contextual details
    a. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
    b. What impact would the naming conventions, proprietary software, and software updates have on later data access?
    c. What other contextual details would you specifically need to document to make your data meaningful to others?
    d. In what form will you capture these details?

3. Data Storage, Backup, Security

    a. Where and on what media will the data from each data source be stored?
    b. How, how often and where will the data from each source be backed up?
    c. How will you manage data security across research assistants/residents on the study for each data source?
    d. How long following the completion of your study will you store the data?

4. Data protection/privacy

    a. How are you addressing any ethical or privacy issues?
    b. What mechanism are you using to identify individual patients?
    c. Who will own any copyright or intellectual property rights to the data from each source?
    d. How will the dataset be licensed if rights exist?
    e. How will the data be associated with a study ID?

5. Policies for reuse of data
    a. How will you create a de-identified copy of the data?

b. Will a new patient consent be required for subsequent re-use of data collected specific to the purpose of this study?
c. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
d. Are there any reasons not to share or re-use data?

6. Policies for access and sharing

a. Will some kind of contribution or fee be charged for subsequent access to this data?
b. What process should be followed to gain future access to your study data?

7. Archiving and preservation

a. What is the long-term strategy for maintaining, curating and archiving the data?
b. What data will be included in an archive?
c. Where and how will it be archived?
d. What other contextual data or other related data will be included in the archive?
e. How long will the data be kept beyond the life of the project?

Case B: Regeneration of Functional Heart Tissue in Rats

Summary of Teaching Points

Module 1: Overview of Research Data Management

- Paper Lab notebook inconsistencies across users
- Lack of synchronization between lab notebook entries and surgical log

Module 2: Types, Formats, and Stages of Data

- Data sources are linked together on an Excel spreadsheet

Module 3: Contextual Details

- Naming conventions for data sets

Module 4: Data Storage, Backup, and Security

- Lack of consistent plan to keep track of slides (in refrigerators) containing tissues and of stages of tissue processing
- Storage issues for large number of optical and electronic images
- Lack of backup for lab notebooks
- Backup plan for data but obviates usefulness of naming conventions

Module 5: Legal and Ethical Issues

- Home grown analysis software ownership and preservation

Module 6: Data Sharing and Re-Use

- Home grown analysis software ownership and preservation

Module 7: Plan for Archiving and Preservation of Data

- None

Research Data Management Case B: Regeneration of Functional Heart Tissue in Rats

The goal of the study is to try to regenerate functional heart tissue in a rat. Unlike other organs and tissues which regenerate themselves, the heart does not have the ability to regenerate, so we intend to regenerate it by delivering stem cells to the heart. The hope is that in generating heart tissue, we generate tissue that is actually functioning and contracting and doing mechanical work.

Two days before we operate on the rat, we take adult stem cells and we incubate them for 24 hours with our marker for cells [fluorescent nanoparticles]. We then put them in a solution and inject them into a tube that has a biological suture in it, so the cells sit down on the outside of the biological suture. We incubate it for 24 hours, and then do the surgery. During the surgery, we open up the thoracic cavity of the rat and create a myocardial infarction by occluding the left anterior descending coronary artery. At this point it is ischemic; we keep it ischemic for 1 hour, not letting any blood flow go through, and then we reperfuse it and let the blood go back. About a minute after that, we put the biological suture with the cells on it through the infarcted region. We then close the rat up and put it back in the cage for a week. We go back a week later, open the rat up again, and use our camera system to acquire images of the heart. We take images with two cameras simultaneously and we'll also have a pressure transducer which syncs automatically with the pictures inside the left ventricle cavity to measure left ventricle pressure. Then we reposition our cameras and take another data set and we usually do that about 4 or 5 times to look at different regions around that infarct.

Then we euthanize the animal. We isolate the heart. We fix the heart in a fixative and then put it in the freezer for about 24 hours. Then we start cutting sections of the heart and putting them onto slides – about 3 sections of the rat heart per slide. We generate about 200 slides per rat heart. At any time, some tissue that was sectioned and on slides may be in one freezer, and some tissue that had not been sectioned yet but was embedded and ready to be sectioned is in another and still other tissue that may be sitting in a container someplace in another freezer. It should be entered into the excel spreadsheet saying what was done and where it is, but that doesn't always happen. Then we stain some slides and then sometime after – anywhere from a day to a couple months - we stain some of them with trichrome. That tells us what tissue is dead. We stain some of them for specific markers in looking to find out exactly where the stem cells are in that cross section. Then we take images on our microscope, which is an epifluorescent microscope and, if we are happy with the staining and the way they look, then we make an appointment to use the confocal microscope which takes much better quality pictures and take those images on the confocal. At the same time, we also look at the data we acquired and use our home-grown custom software to track particles on the surface of the heart to see how far and how fast those particles are moving. The software was written in C and MATLAB (C runs the code faster, but MATLAB is easier to work with; usually we develop the code in MATLAB and then convert it to C so that it runs faster). We use this software to analyze

the optical images of the heart. That tells us what the function is like in that region of the heart. We do that for several heartbeats in different data sets. And then we save that data. That is everything we do for one heart.

Data sets:

1) We have the optical images after the first surgery to insert the cells – on average for one experiment we probably have about 10,000 images. We acquire images of the heart at about 250 frames/second. We acquire 4 seconds worth of data, so we have 1,000 images for each data set. The images are initially stored on the hard drive of the acquisition computer, then are transferred to a Drobo backup system and the hard drive of a network computer that is backed up by the institution.

2) The second data set is where we measure the left ventricle pressure at the same time we are acquiring those images so that we know that image # 127 correlates with the pressure at time point # 127 milliseconds. For this measure we use an analog to digital (A/D) board and a Millar pressure transducer. Both the camera and the acquisition system are computer controlled to synchronize them. These data are stored in the same way as the optical images, although they are separate files.

3) Electronic data sets are used to acquire images from the different stained tissue sections after the second surgery. We may stain on average 4-5 different markers and we will have different data sets for the different stains. So we will have images taken from the epifluorescent scope, and, usually in a cross-section of heart, we may take some high resolution images in zoomed in regions and some low magnification images. On average for each section, we take about 20 pictures with the epifluorescent and then with the confocal, we probably take about another 20; if we take a z-stack with the confocal, that can be an additional 200 images. They are both taking the same images except one is a much better resolution than the other. Long term storage for these datasets is on the Drobo and DVD backups.

Our naming convention is that we name our files EXP (for experiment) and then usually 4- or 5-digit codes like 2001, and then we have several data sets so it is DS # and then it could be image 1. Then we need to link these data sources together via an excel spreadsheet. The sections are all linked by the same experiment number. They are linked with the digital images just based on what number section they are.

Multiple research staff may be analyzing the same heart, and one person will be doing the mechanical function of the heart, one will be doing the trichrome staining, another will be doing the actinin staining and maybe another will be doing the imaging. The data sets should all be linked in the excel spreadsheet. There could easily be up to 10 people involved in data analysis, and we have not yet found a good way to link all the data. We have an excel spreadsheet basically, and it says in DS1 – the mechanical function in this area was xyz, in DS2 it was this. In DS1 the tissue section showed this, and we try to link them all up together, but the tissue sections are on conventional microscope slides that are stored someplace. Even that – the location of where they are stored is a problem. We

have the usual places where we store things but we have 3 or 4 freezers and if it is not in 1, we look in 2, and so on. The slide box is labeled with the experiment number and the individual slides are labeled with the slide number & experiment number.

The types of data we use are mostly images and numeric measures in addition to the lab notebook which may have some observational notes. Some of it is number-crunching but a lot of it is images.

The content of a lab book relates to a particular experiment and is used by all staff working on that experiment. There is a format they are all supposed to follow, which they don't always do. There could be on average 5-6 people using the notebook. The paper lab notebook basically performs the function of being an index into the actual datasets and it should record all the information the PI specifies. We also have a paper surgical log that is kept with the animal and whatever project staff writes down in that surgical log should be transferred into the lab notebook – so it has to be in 2 places. It has to be down there in case there is a problem with the animal, but the PI also needs it in the lab notebook to be able to write papers. The older lab notebooks are in the PI's office, but the ones that are currently in use are in the lab. Older Lab notebooks are only in the PI's office of lab with no backup. The lab notebook has to be in pen on specific paper because this paper is supposed to be good for 100 years.

We backup the data sets on an external hard drive someplace. The optical and electronic images are both backed up. The current backup system we are using truncates the data set name to 6 digits then puts a tilde sign and number starting from 001.

The files are not password protected or anything. The lab notebook is either in the PI's office or (most often) in the lab, which has key card access (although it is an open floor plan).

Module 1: (Overview module) discussion question:

What issues need to be addressed on this project related to the 7 segments of the data management plan components?

Discussion Questions for Other Modules:

1. Types of data

    a. What types of data are being collected for this study?
    b. How will you ensure all research staff used the same data sources and data definitions?
    c. What would be needed in a data management plan to describe use of novel equipment?
    d. What needs to be in the plan related to the data capture for the various data sets?
    e. What analytical methods and mechanisms will be applied to your data either prior to or post integration?
    f. What type of outcome data will be generated?

2. File Formats and Contextual details

    a. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
    b. What impact would the naming conventions and the use of homegrown software have on later data access?
    c. What other contextual details would you specifically need to document to make your data meaningful to others?
    d. In what form will you capture these details?

3. Data Storage, Backup, Security

    a. Where and on what media will the data from each data source be stored?
    b. How, how often and where will the data from each source be backed up?
    c. How will you manage data security across research staff on the study for each data source?
    d. How long following the completion of your study will you store the data?

4. Data protection/privacy

    a. How are you addressing any ethical or privacy issues?
    b. What mechanism are you using to identify individual animals or hearts?
    c. Who will own any copyright or intellectual property rights to the data from each source?
    d. How will the dataset be licensed if rights exist?
    e. How will the data be associated with a study ID?

5. Policies for reuse of data

    a. Will you need to create a de-identified copy of the data?
    b. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
    c. Are there any reasons not to share or re-use data?

6. Policies for access and sharing

    a. Will some kind of contribution or fee be charged for subsequent access to this data?
    b. What process should be followed to gain future access to your study data?

7. Archiving and preservation

    a. What is the long-term strategy for maintaining, curating and archiving the data?
    b. What data will be included in an archive?
    c. Where and how will it be archived?
    d. What other contextual data or other related data will be included in the archive?
    e. How long will the data be kept beyond the life of the project?

Case C: Improving End-of-Life Care for African Americans

Summary of Teaching Points

Module 1: Overview of Research Data Management

- Lack of planning for how to transfer data to and from contracted analyst at other university

Module 2: Types, Formats, and Storage of Data

- Audiotapes, Microsoft Word document

Module 3: Contextual details

- No naming conventions for tape-recorded data
- Transcribed data reviewed for accuracy

Module 4: Data Storage, Backup and Security

- Lack of plan for backup of transcribed data and outcome data

Module 5: Legal and Ethical Issues

- Informed consent needed from participants
- De-identification of participants needed

Module 6: Data Sharing and Re-Use

- Further de-identification of transcript subsets for re-use in methodology course

Module 7: Plan for archiving and preservation of data

- No plan for preservation of source or outcome data at PI's site other than published manuscript

Case C: Improving End-of-Life Care for African Americans[1]

An MD applied for grant funding to do a qualitative study focusing on how to improve physician communication with African Americans (AA) and their relatives when their patients were receiving end-of-life care.

This qualitative study was conducted to expand knowledge about AA experiences and opinions about end-of-life care. Multiple-meeting focus groups were held to build trust and allow time for full participation. Following a review by a Community Advisory Board (CAB), protocols were approved by the University's Institutional Review Board. Participants were AA adults who had experienced at least one death of a significant other or family member. Convenience sampling by staff and CAB members was used to recruit participants, and flyers were distributed at neighborhood activities. Participants were screened for eligibility and assigned to one of two focus groups. Focus group 1, which met for four sessions, was comprised of AAs with family members who had died at home. Focus group 2 met for three sessions and included AAs with family members who had died in the hospital. An average of five individuals attended each session. Three participants worked in health care, and their observations reflected experiences with a dying family member, as well as experiences with caring for terminally ill AA patients.

Data collection All participants gave informed consent. An open-ended interview script stimulated discussion about (1) positive and negative experiences of participants related to end-of-life care in the hospital or at home, (2) preferences for treatment by health care providers, (3) communication issues, and (4) end-of-life decision making pertaining to living wills and advance directives. An AA member of the project staff moderated the focus groups.

Each session was audio-taped taped. Unlabeled tapes were mailed to a transcriptionist in their plastic cases which were labeled. During the mailing process the package was damaged and the plastic tape cases broke and were no longer associated with the tapes for which the cases had been labeled. The tapes, however, were not damaged. The transcriptionist transcribed the tapes and the transcripts were sent back to the project team for identification of which focus group and which session should be used to identify each transcript. Focus Group Participants' comments were identified on the transcript by either Miss, Mrs. or Mr. plus the first initial of their first name. The transcripts were also reviewed for accuracy by the project team.

Data analysis Transcripts were reviewed for themes through a continuous process of text data segment comparison based on qualitative research techniques. After reading the transcripts several times, a codebook was developed defining themes and subthemes, and a numeric theme code was assigned to each particular category of text responses. Microsoft Word was used to create transcript tables of participant responses which then could be sorted by theme code. Participants' responses were coded and sorted accordingly into differing categories which were then summarized to capture the richness and range of data within each theme code. The analysis was systematic and involved triangulation of data from the two focus group sources. Within-focus group set analyses were performed, as well as cross-focus group set analyses to develop a set of themes/recommendations for how end-of-life care communications might be conducted to improve the process for all concerned.

Resulting Data: In a subsequent publication[1], the results were published as follows: Analysis of the transcripts revealed five major theme groupings. These groupings contained text data related to:

1. Communicating about dying and end-of-life care

2. Choice about dying at home or in the hospital

3. Dying in the hospital

4. Dying at home

5. Other end-of-life care issues

Additionally the implications for clinical care were summarized as follows:

- Be mindful of the diversity of preferences and needs within any population subgroup
- Recognize that many AAs have very strong religious and spiritual beliefs about dying and that their words often reflect that the patient is preparing to leave his or her earthly home
- Empower dying AAs and their family members by speaking respectfully, using lay terminology, and checking for understanding. Encourage the patient to be the primary decision maker and ensure that the dying person is not infantilized.
- Determine whether the dying person and/or caretaker has adequate assistance. Since awareness of home and hospice services is low, facilitate getting necessary support and resources, including connections with social services.
- Encourage patients to decide how the family should be informed about prognosis and provide assistance in telling the family if requested.
- Determine in advance who the primary family contact is and where to contact him or her in the final hours if the patient is hospitalized. If possible, ensure that the family has the opportunity to spend the last hours with the patient. The "gathering of the family" is very important during this phase of life.
- For patients dying in the hospital, treat patients the way you want to be treated with nurturing, compassion, dignity, love, touch, and careful listening. Diligent monitoring of the patient's medical status, needs, and cleanliness is imperative.

The tapes were eventually destroyed and the transcripts and other files generated during the analysis remained with the analyst who was not part of the project team and was affiliated with another medical school. The analyst was very involved with the drafting of the publication. Excerpts from the transcripts were later re-used as examples for a qualitative analysis class taught by the analyst; however, for the reuse, all participant IDs were changed to P1, P2, etc.

---

[1]Case loosely modeled after the study described in *End-of-Life Care and African Americans: Voices from the Community*, CAROLYN JENKINS, Dr.P.H., F.A.A.N., NANCY LAPELLE, Ph.D., JANE G. ZAPKA, Sc.D., and JEROME E. KURENT, M.D., M.P.H., JOURNAL OF PALLIATIVE MEDICINE, Volume 8, Number 3, 2005, © Mary Ann Liebert, Inc.

Module 1 (Overview module) discussion question:

What issues need to be addressed on this project related to the 7 segments of the data management plan components?

Discussion Questions for Other Modules:

1. Types of data

    a. What types of data are being collected for this study?
    b. How will you ensure all research staff used the same data sources and data definitions?
    c. What needs to be in the plan related to the data capture for the various data sets?
    d. What analytical methods and mechanisms will be applied to your data either prior to or post integration?
    e. What type of outcome data will be generated?

2. File Formats and Contextual details

    a. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
    b. What impact would the naming conventions have on later data access?
    c. What other contextual details would you specifically need to document to make your data meaningful to others?
    d. In what form will you capture these details?

3. Data Storage, Backup, Security

    a. Where and on what media will the data from each data source be stored?
    b. How will you manage data security across research staff and transcriptionist on the study for each data source?
    c. How long following the completion of your study will you store the data?

4. Data protection/privacy

    a. How are you addressing any ethical or privacy issues?
    b. Who will own any copyright or intellectual property rights to the data from each source?
    c. How will the data be associated with a study ID?

5. Policies for reuse of data

    a. Will you need to create a de-identified copy of the data?
    b. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
    c. Are there any reasons not to share or re-use data?

6. Policies for access and sharing

    a. Will some kind of contribution or fee be charged for subsequent access to this data?
    b. What process should be followed to gain future access to your study data?

7. Archiving and preservation

    a. What is the long-term strategy for maintaining, curating and archiving the data?
    b. What data will be included in an archive?
    c. Where and how will it be archived?
    d. What other contextual data or other related data will be included in the archive?
    e. How long will the data be kept beyond the life of the project?

Case D: Characterizing a Component of a Rocket Engine used to Control Satellites in Orbit

Summary of Teaching Points

Module 1: Overview of Research Data Management

- Paper Lab notebook inconsistencies across users

Module 2: Types, Formats & Stages of Data

- None

Module 3: Contextual Details

- Lack of documentation for MATLAB code for ease of use by future users
- No standards for homegrown analysis software documentation

Module 4: Data Storage, Backup and Security

- ITAR restrictions on foreign nationals having access to research on cathode developed by a private company with Air Force funding
- Equipment and lab notebooks kept in locked cabinet
- Storage of the MATLAB code
- No backup plan for the lab notebooks or archived MATLAB code CDs
- Source data is retained on PI's computer and automatically backed up by the institution

Module 5: Legal and Ethical Issues

- Ownership of the MATLAB code used to display results

Module 6: Data Sharing and Re-Use

- Restrictions on providing details in publications related to NASA cathode design making reproducibility by other researchers difficult

Module 7: Plan for Archiving and Preservation of Data

- None

Research Data Management Case D: Characterizing a Component of a Rocket Engine used to Control Satellites in Orbit

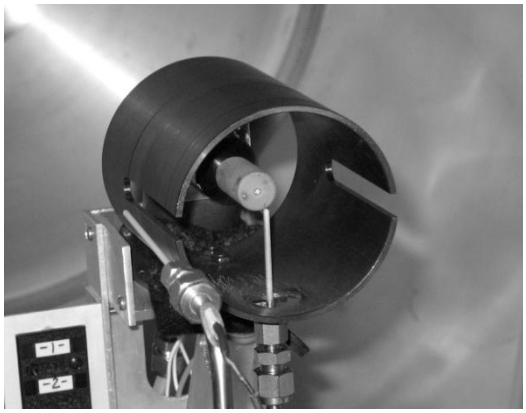


Fig. 1 50" x 72" vacuum chamber used for cathode research (WPI Higgins Laboratory)

Fig.2. Hollow cathode and surrounding anode installed in vacuum chamber. Also visible are Langmuir and emissive probes (small white rods).

Data Issues Addressed (Modules):

- Data Management/Ownership
- Data Life Cycles
- Data Storage
- Data Security
- Data Privacy/Restrictions

Suggested Curriculum Levels:

- Undergraduate, years 3-4
- Graduate

Suggested student populations:

- Engineering
- Lab Experimentation – Physical Sciences

Scenario:

A faculty researcher in Aerospace Engineering studies electric propulsion for spacecraft control, in other words a type of rocket engine that uses electricity to ionize and accelerate a gas to produce thrust. He is concerned about data storage and security in his lab and is looking for a

standard protocol that could be used by all and would comply with any data storage and security requirements imposed by his research sponsors.

The goal of one of his current projects is to study and characterize a component of an electric thruster being used by NASA, the Air Force, and private companies to control satellites in orbit. This work enables researchers to build more robust thrusters that will have a longer service life than current models, thus enabling longer and more ambitious space missions. The students in the lab have been experimenting on a particular engine component, called a "hollow cathode," to characterize the plasma it generates. This data will help researchers understand where energetic ions are produced that erode surfaces and limit the cathode lifetime. They perform experiments using Langmuir and emissive probes to collect data from two different cathodes tested in a vacuum chamber.

The two cathodes used in the lab have different restrictions on their use. The first is from a private company, was developed with Air Force funding, and is the same model as a unit which has been used operationally in orbit. Because it is identical to flight hardware, work with this cathode must comply with International Traffic in Arms Regulations (ITAR) requiring that no foreign nationals have access to any aspect of the research. ITAR establishes strict controls on the use and dissemination of information related to defense articles. Some equipment and lab notebooks are kept locked up in order to comply.

The second cathode is from NASA, is a laboratory use model (i.e. not "flight hardware"), and is not subject to the same ITAR restrictions. However, NASA does place some restrictions on publication related to the cathode design, such as specific geometry and dimensions. This makes it difficult to publish this type of research and data since the experiment would not be reproducible by other researchers unless they have access to the same devices.

Raw data is generated from the cathodes during experiments and is downloaded onto a laptop. The experimental conditions are recorded in the laboratory notebook. Students review the data and discard any that is not useful. For useful data, the students produce code using MATLAB and create appropriate graphs and charts displaying the useful data points. They share findings with the researcher so that all can discuss and determine whether there are reportable results that will be useful to the aerospace community. The MATLAB code is the intellectual property of the student and faculty advisor who create it. If code is written primarily by a student, then the advisor will usually try to contact the student (who may have graduated) for permission before distributing it to another research group. Permission is rarely granted for sharing this code outside the research group.

Over the past two years the researcher has had four graduate students work on this project. Two of the students have graduated and are no longer part of the team.

When students graduate they are required to provide the researcher with a CD containing the MATLAB code used for any analysis they completed in the lab. These CDs are then stored in the researchers office in case of a future need. This code is not always sufficiently documented so that it could be easily used in the future. Some students do a better job than others at describing how their code works and the researcher is interested in learning about how to standardize this practice, although he has never had a problem in the past.

Laboratory notebooks are generally kept in the laboratory. Notebooks which include any information on projects which are ITAR restricted are kept in a secure location, either in the researcher's office or in a locked storage cabinet located in a lab. The archive CDs are stored in

the researcher's office and there is no backup plan for either the CDs or the notebooks. There is also a minor concern over the fact that laboratory notebook entries are lacking some information and may need to be more standardized. In some cases there are no initials indicating who completed an entry in the notebook. Other descriptive elements may be missing as well.

The only data that is retained is that which has been used to generate results presented in either a thesis or a publication. This data is stored on the researcher's computer which is automatically backed up nightly through his institution.

Module 1 (Overview module) discussion question: What issues need to be addressed on this project related to the 7 segments of the data management plan components?

Discussion Questions for Other Modules:

1. Types of data

    a. What types of data are being collected for this study?
    b. How will the data be captured?
    c. How will you ensure all research staff used the same data sources and data definitions?
    d. What needs to be in the plan related to the data captured from testing the two different cathodes?
    e. What analytical methods and mechanisms will be applied to your data either prior to or post integration?
    f. What type of outcome data will be generated?

2. File Formats and Contextual details

    a. What file formats and naming conventions will be used for the separate data sources and the MATLAB code? What impact would the naming conventions have on later data access?
    b. What other contextual details would you specifically need to document to make your data meaningful to others?
    c. In what form will you capture these details?

3. Data Storage, Backup, Security

    a. Where and on what media will the data from each data source be stored?
    b. How will you manage data security across research staff on the study for each data source?
    c. What concerns are there regarding the security of the data that is kept on a CD, laptop, or in the lab notebook?
    d. How long following the completion of your study will you store the data?

4. Data protection/privacy

    a. How are you addressing any ethical or privacy issues?
    b. Who will own any copyright or intellectual property rights to the data from each source?
    c. How will the data be associated with a study ID?

5. Policies for reuse of data

    a. Will you need to create a de-identified copy of the data?
    b. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
    c. Are there any reasons not to share or re-use data?

6. Policies for access and sharing

    a. Will some kind of contribution or fee be charged for subsequent access to this data?
    b. What process should be followed to gain future access to your study data?

7. Archiving and preservation

    a. What is the long-term strategy for maintaining, curating and archiving the data?
    b. What data will be included in an archive?
    c. Where and how will it be archived?
    d. What other contextual data or other related data will be included in the archive?
    e. How long will the data be kept beyond the life of the project?

Course Module 5 (Fully developed content)

Legal and Ethical Considerations for Research Data

By participating fully in this class, student will be able to:

1. Explain ownership considerations related to data sharing
2. Explain and evaluate potential legal issues connected to one's data; intellectual property, copyright claims, licenses needed for use, monetary charges for data
3. Explain ethical considerations related to data sharing
4. Understand privacy levels for research data as required by potential funding agencies
5. Recognize the importance of privacy with some forms of research data (HIPAA)
6. Understand the importance of removing key personal identifiers to facilitate confidentiality
7. Understand the need for data attribution and citation.

Section 1: Ownership

Who owns the data?

This question pertains to who has the legal rights to the data, who can retain the data after the completion of the project, and whether the PI (Principal Investigator) has a right to transfer data between institutions.

Ownership of the data really depends on who funds the research. Funders sponsor research for a variety of reasons:

- Government agencies fund research to improve the general health and welfare of society
- Philanthropic organizations are interested in advancing particular causes
- Private funders are interested in profits, along with benefits to society

These different reasons often determine who claims ownership of research data.

For federally funded grants:

In most cases for federally funded research, the government gives the research institution the right to use data collected with public funds as an incentive to put research to use for the common good (the Bayh-Dole Act). Thus the research institution owns the data but allows the principal investigator on the grant to be the steward of the data. The PI may control the course, publication, and copyright of any research, subject to institutional review. Graduate students, postdocs, or faculty involved in performing research on a particular grant would therefore be wrong to assume that they own the data that they are collecting. The PI takes responsibility for the collection, recording, storage, retention and disposal of data.

Data and lab notebooks collected by undergraduate and graduate students and research fellows for a research project belong to the grantee institution. Students should not take the data with them when they leave the institution unless they have made appropriate arrangements with the project PI. Retaining copies of data might also be allowed, with permission.

When the PI faculty member leaves the grantee institution, they must negotiate with the institution to keep their grants and data. Many universities have offices and policies in place to ensure that such a transfer of data respects both the rights of the researcher and those of the institution(s).

Is it a grant or a contract?

With government funding, researchers should also distinguish between grants and contracts. Under grants, researchers must carry out the research and submit reports, but control of the data remains with the institution that received the funds.

With contracts, the researcher is required to deliver a product or service, which is then usually controlled by the government. If your research is supported with government funds, make sure you know whether it is a grant or a contract. This is a significant difference that could determine who can publish and use your data.

Private funding companies

Private funders seek to retain the right for commercial use of the data.

Philanthropic organizations

Their policies can vary. Depending on their interests, they may retain or give away ownership rights.

As you see, ownership claims do vary from one funder to another. Therefore it is crucial that researchers be aware of their obligations to their funders before they begin collecting data.

Reading:

Guidelines for Responsible Data Management in Scientific Research
http://ori.hhs.gov/education/products/clinicaltools/data.pdf
pgs. 6-8

Section 2: Data and Intellectual Property

When preparing for a research project involving data, be sure to evaluate all the legal issues: intellectual property, copyright claims, licenses needed for use, monetary charges for data and other intellectual property issues.

Also consider the different types of outputs within a research project:

- Process

- Datasets
- Publications
- Software code

Intellectual Property Overview

According to the World Intellectual Property Organization (WIPO), intellectual property is defined as, "creations of the mind: inventions, literary and artistic works, and symbols, names, images, and designs used in commerce." So in other words, they are basically intangible assets comprised of knowledge and ideas. Intellectual property generated in an academic setting usually involves copyrights, trade secrets, and patents.

The creation of intellectual property is one of the expected outcomes of research conducted at universities. It benefits both the university and society to facilitate the development of these discoveries and ideas as well as to assure their availability to the public. With these goals in mind, universities develop policies and procedures relating to the ownership, use, management, and compensation for intellectual properties created with their resources. Because Intellectual Property Policies vary by institution, be sure to familiarize yourself with your institution's policies. A few sample policies are listed here:

- http://www.umassmed.edu/otm/ippolicy.aspx
- http://www.provost.duke.edu/pdfs/intelProp.pdf
- http://www.wpi.edu/offices/policies/intell.html

Data can be licensed so you need to think about the issue from both sides; i.e., as a creator of data and as a user of others' data.

For your research project, you will need to articulate how you will be providing permissions or licensing to your data or copyrighted works from your research project. Factors you may want to consider are:

- Attribution
- Notification regarding its use
- Redistribution
- Quality control
- Risk

In cases where government funded research data is protected by intellectual property rights, rights holders should facilitate data access for the benefit of public research. As the National Science Foundation (NSF) states:

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing."

The NSF also now requires a data management plan and addresses intellectual policies on their web site: http://www.nsf.gov/od/ogc/intelprop.jsp

For further information, see the NSF's Frequently Asked Questions (FAQ) site on data sharing.

Copyright

A copyrightable work is an original creative work set in a tangible format that is covered by the copyright laws of the United States or other countries. Copyright protection is available for most literary, musical, dramatic, photographic and other types of creative works--including research articles, research monographs, textbooks, student theses and dissertations, still images, computer software, teaching materials, multimedia works, proposals, and research reports. Copyright is "format blind" – that is, print and digital works are eligible for copyright protection; content on the Internet may be protected by copyright.

Copyright ownership is secured automatically when an original creative work is fixed in a tangible format; the © is no longer required. Ownership may belong to the author/creator or their employer.

The copyright owner has the exclusive rights over the work to reproduce it, distribute copies, create a derivative work, perform or display the work publicly. Ownership rights may be transferred in whole or in part; in the past, authors often transferred all of their rights to their works to journal and book publishers.

Not all works are eligible for copyright protection: original works that are not fixed in a tangible format; titles, names, slogans; ideas, facts, data; lists of contents or ingredients; works in the public domain. However, trademark or patent laws may apply in some cases.

Patents

The United States Patent and Trademark Office (USPTO) defines the patent for an invention as, "the grant of a property right to the inventor" to "exclude others from making, using, offering for sale, or selling the invention in the United States or importing the invention into the United States." Patents are usually granted for twenty years but the term may be extended. U.S. patents are effective only within the United States, U.S. territories, and U.S. possessions. Because a patent may be challenged at any time during its twenty year term, it is important to preserve the related data for at least the term of the patent. Good data management practices could also provide more efficient problem resolution if a patent official discovers any data irregularities while evaluating a patent application.

Whether the source of the funding is federal or private, there are likely to be certain obligations regarding intellectual property, especially with relation to inventions and patents. It may be mandated by law, by contract, or both. Prior to 1980, inventions that resulted from federally funded research grants and contracts were under the control of the federal government. However, since the passage of the Bayh-Dole Act in 1980, universities, small businesses, and non-profits may choose to retain title to inventions developed with federal funding. This gives universities greater incentive to practice, for example, data mining which in turn may lead to inventions and the patent applications needed to protect those inventions.

Trade Secrets

Trade secrets are generally confidential commercial information such as formulas, manufacturing processes, or compilations of information which are automatically protected without any formal

registration procedures (e.g., formula for Coca Cola©). Trade secrets are generally protected under state law. Keep in mind that data about or from companies might contain proprietary data which is not accessible for research purposes.

Open Source Software

Open Source software is computer software often developed in a collaborative manner. The source code is made widely available through a type of license that allows users to freely, modify, improve, and redistribute the software as long as they agree to the conditions specified in the license provided. Before agreeing to an open source software license, make sure that:

- Your funder/sponsor agrees to the conditions of use
- The conditions do not adversely impact your intellectual property rights

Research Datasets and Databases

The U.S. Federal Government's Office of Management and Budget Circular A-110 (36.d.2.i Property Standards; Intangible property; definition) states:

Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples). Research data also do not include: (A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and (B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.

An important point to consider is that in the United States, while data and facts cannot be copyrighted, creative expressions of data, such as a chart or a table in a publication ARE copyrightable. In addition, be aware that in certain foreign jurisdictions such as the European Union, database compilations including factual data ARE protected by law.

Databases are generally protected by copyright law and are referred to as “compilations.” The U.S. Copyright Act defines a compilation as a “collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship.”

The individual facts or data contained within the database may or may not be protected by copyright; however, the selection and/or arrangement of the facts or data as a whole will be protected by copyright if it contains enough creative, original expression.

With only limited protection through copyright law, database developers generally protect their databases by using a legal contract, such as a license, so that users must comply with wishes of the copyright owner as to how that data may be accessed and used.

ACTIVITY

Read scenario "Who Owns Research Data?" (Case scenario of graduate student wanting to take data, case study from Columbia University Responsible Conduct of Research Data Acquisition and Management), discuss five follow-up questions.

http://ori.dhhs.gov/education/products/columbia_wbt/rcr_data/case/index.html#2

Reading:

"Constructing Access Permissions", University of Oregon Libraries:
http://libweb.uoregon.edu/datamanagement/sharingdata.html#three

Section 3: Ethics and Data

Any research institution (university, hospital, private research company, and so on) that accepts federal funding is required by law to have policies in place to oversee its research programs. These policies include monitoring conflicts of interest, reporting misconduct, and ensuring safety regulations are followed, as well the establishment of standing committees to review human (Institutional Review Board) and animal (Institutional Animal Care and Use Committee) research protocols.

The purpose of an Institutional Review Board (IRB) is to protect the rights and welfare of those individuals who contribute to the research process by participating as subjects. The IRB also protects the institution and the researcher by ensuring that those individuals considering being part of a research study are adequately informed before consenting to participate, and that participants are not exposed to excessive risk.

In the context of data management the IRB has three roles. First, since funders often now ask to see data management plans, members of the IRB look more closely at these plans to see if adequate thought has been given to the plan and if what is written is feasible (cost, infrastructure, staffing). Second, the IRB reviews data collection forms to limit the amount of personal identifiable information that is being collected. Third, the IRB reviews the research protocol to see how the data will be safeguarded. This includes documenting who will have access to the data collected, and under what conditions – sometimes called the privacy or confidentiality rules. These rules need to consider who will have access to the data technically, physically and for administrative purposes.

There are federal and state rules and regulations regarding data security for specific types of data. For instance, personal identifiable data, such as names and social security numbers, are protected by many state and federal laws. At the federal level, health data are protected by the federal Health Insurance Portability and Accountability Act (HIPAA), student data are protected by the federal Family Education Rights and Privacy Act (FERPA), and financial data are protected by the federal Financial Services Modernization Act (FSMA).

As researchers work to collect and analyze data they must ask themselves if each piece of data is necessary to address the original research question or hypothesis and if the data element in combination with other data could identify an individual. For example, age alone may not identify a person, but age in conjunction with zip code and medical condition may lead to identification. To protect confidentiality in these instances, researchers should not collect the data at all, or if it is crucial, should substitute the actual data with codes known only to the primary researcher. The HIPAA Privacy Rules outline 18 data elements that need to be coded or removed (http://healthcare.partners.org/phsirb/deidinfo.htm).

Privacy levels required by funding agencies and publishers

Each funding agency and publisher has guidelines for maintaining privacy regarding human and animal subjects, as exemplified in this guideline from the National Institutes of Health (NIH):

"Data should be redacted to strip all individual identifiers, and effective strategies should be adopted to minimize risk of disclosing a participant's identity. Options to protect privacy include: withholding part of the data, statistically altering the data in ways that will not compromise secondary analyses, requiring researchers who seek data to commit to protect privacy and confidentiality, and providing data access in a controlled site, sometimes referred to as a data enclave. Some investigators use hybrid methods, releasing a redacted dataset for general use but providing access to more sensitive data through a user contract or data enclave. In most instances, sharing data is possible without compromising participant confidentiality and privacy."

(NIH's Office of Extramural Research:
http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#923)

The National Science Foundation requires that the following question be addressed in all grant proposals' data management plans:

"What will be the policies for data sharing and public access (including provisions for protection of privacy, confidentiality, security, intellectual property rights and other rights as appropriate)?"

Using Data Created by Others

When making a request to use another's data, be specific. Are you looking for an entire dataset that includes multiple variables or are you looking for a subset of this data? Are you looking for data from a specific time frame? In a specific geographic area? Knowing and communicating to the data creator what the "boundaries" are of the data you want to access facilitates the sharing process.

Section 4: Citation / Attribution of Data

Acknowledgement of the use of someone else's information or work is a long-accepted practice in scholarly communication. This attribution is usually accomplished via a "citation": when you publish a paper or do a presentation that makes use of someone else's information, you include in that paper or presentation a reference to the work of that other person or group. It is important to cite not only the literature consulted but also the data files used, including your own.

Citing data files in publications based on those data serves several purposes:

- Provides appropriate credit to the data producers and publishers
- Enables other researchers to access the data for their own use or to replicate research findings
- Assists in measuring the impact of a dataset by tracking references to it in the scientific literature
- Helps data producers know how their data is being utilized

Data citation is evolving and there is currently no acknowledged standard on how to cite a dataset or construct a data citation. However, several organizations and data stewards have developed their own practices, and international groups are creating formal guidelines for the scientific community. The following elements are generally considered the core elements of a citation:

- Author(s) – the creators of the data; can be one or more people or organizations
- Title – the title of the data set
- Version – the exact version or edition of the dataset used
- Release Date – the date when the dataset was published or released
- Publisher/Archive – the data center or repository that is archiving and distributing the data
- Identifier/Locator – URL or other locator for the data; a persistent URL such as a DOI or a Handle is preferred
- Access Date – the date when the online dataset was accessed

Example of a data citation (following the International Polar Year and Data Information Service format):

Bockheim, J., Cline, D. 2003. CLPX-Ground: ISA snow pit measurements.
Version 2.4, Sept. 2003. Boulder, Colorado, USA: National Snow and Ice Data Center/World Data Center for Glaciology. http://nisdc.org/data/arCss006.html
Data set accessed 2008-05-14.

(Citation crafted from data citation examples from the International Polar Year Data and Information Service http://ipydis.org/data/citations.html)

Repositories often provide guidance on how to cite their data sets. Here are some specific guidelines and practices for data citation:

- International Polar Year and Data Information Service: "How to Cite a Data Set" http://ipydis.org/data/citations.html

- Dryad, an international repository of data underlying peer-reviewed articles in the basic and applied biosciences: "How should I cite data from Dryad?" http://datadryad.org/using#howCite

- Dataverse Network Project: "Data Citation Standard" http://thedata.org/citation

- DataCite consortium: "Why Cite Data?" http://datacite.org/whycitedata and metadata schema of core elements of a data citation http://schema.datacite.org/

- Federation of Earth Science Information Partners: "Interagency Data Stewardship/Citations/Provider Guidelines" http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

Readings:

Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4), http://www.dlib.org/dlib/march07/altman/03altman.html

Ball, A. & Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

International Polar Year Data and Information Service: How to Cite a Data Set http://ipydis.org/data/citations.html

ACTIVITY:

Have students identify the components of the following data citation:

Spencer, R., Roseman, I. 2007. *COLSP/RMS snow swath 1km V005*, Oct. 2006–Apr. 2007. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-08-24 at http://nsidc.org/data/colsp29v5.html

(Fictional citation based on data citation examples from the International Polar Year Data and Information Service http://ipydis.org/data/citations.html)

Assessment:

1. Have students read excerpt of research data management case A or C and respond to questions.

2. Read and discuss commentary "Henrietta's Dance" http://www.jhu.edu/jhumag/0400web/01.html or view "Henrietta Lack-CBS Sunday Morning" http://www.youtube.com/watch?v=wRrNjHYxP_o&feature=related

Excerpt from Research Data Management Case A for use in Module 5

Dr. X wrote a 5 page proposal for funding for a study to use a novel monitor with proprietary software to assess patient outcomes 2 years after orthopedic implant surgery. This prospective longitudinal study would determine the rate of sub-optimal outcomes based on specialized analysis using the proprietary software that accompanied the monitor. The study was funded and the research resident working with the PI prepared the IRB application that received approval. With a clearly defined research hypothesis, innovative monitor technology, and IRB application and consent form complete, the goal was to collect the same measures over 3 years.

The same patients were followed for three years so it required tracking them down to have them come in to allow for collection of data via multiple sources: patient surveys, accelerometer measurements, and surgeon notes from the physical exams. We had HIPAA authorization to use the patient's name, Medical record #, and telephone/address to contact them for follow-up. However, the data base was organized by unique study ID assigned to each patient.

The study was complex due to the need to collect and integrate data from these three different sources:

1) Patient-generated data regarding their demographics and their symptoms, the amount of pain and disability. Patients filled out a hand-developed paper survey at baseline and annually for 3 years. Survey data were entered by various people into an excel spreadsheet and the source documents were stored in multiple locations. Eventually the patient surveys were moved onto a direct computer data entry system, and the data were captured in survey software that could be downloaded into a spreadsheet/data file for analysis.

2) The second source was measurements from an accelerometer that did 24 hour tracing of patients' steps and walking rate annually. This novel monitor came with proprietary analytic software that was on a lab PC originally. It was a proprietary software package that could be loaded only on one computer and it had to be handed off as residents changed. We bought another monitor software license to get it off the original PC because the monitor analytic data were housed there and we then put it on a laptop.

3) The third source was a surgeon note in the EMR. Residents read the charts every month related to patients in the study to identify any follow-up MD office visits and to extract physical exam measures which were inserted into a structured database with data definitions for each measure.

The data from these multiple sources needed to be integrated for a biostatistician to apply longitudinal modeling software. ACCES was the final data base and was used to house the total data set and integrate data (through a flat file) from all the sources. Data sub-sets were imported to STATA software for particular analyses, as needed. Data were stored

on a server solely for research that is password protected, backed up nightly, and protected by institutional firewalls, etc. (not on a computer). Security measures such as passwords, limited access, firewall, etc. were used to safeguard the data.

Data protection/privacy

a) What sorts of privacy conditions might a funder require for the data collected in this study?
b) What ethical or privacy issues in this study relate to sharing data? How can they be resolved?
c) What mechanisms were used to identify individual patients and maintain privacy? Would these need to be changed to preserve confidentiality during re-use?
d) What issues might arise related to copyright or intellectual property rights to the data from each source or outcome data?
e) How might the dataset be licensed or access fees charged to re-users if rights exist?

Research Data Management Case B – Module 5 Excerpt

The goal of the study is to try to regenerate functional heart tissue in a rat by delivering stem cells to the heart.

Two days before we operate on the rat, we take adult stem cells and incubate them for 24 hours with our marker for cells [fluorescent nanoparticles]. We then put them in a solution and inject them into a tube that has a biological suture in it, so the cells sit down on the outside of the biological suture. We incubate it for 24 hours, and then do the surgery. During the surgery, we open up the thoracic cavity of the rat and create a myocardial infarction by occluding the left anterior descending coronary artery. At this point it is ischemic; we keep it ischemic for 1 hour, not letting any blood flow go through, and then we reprefuse it and let the blood go back. About a minute after that, we put the biological suture with the cells on it through the infarcted region. We then close the rat up and put it back in the cage for a week. We go back a week later, open the rat up again, and use our camera system to acquire images of the heart. We take images with two cameras simultaneously and we'll also have a pressure transducer which syncs automatically with the pictures inside the left ventricle cavity to measure left ventricle pressure. Then we reposition our cameras and take another data set and we usually do that about 4 or 5 times to look at different regions around that infarct. Then we euthanize the animal, cut sections of the heart, and put them onto slides. We stain some of them for specific markers in looking to find out exactly where the stem cells are in that cross section and take additional images of these.

In addition to stored images of the living heart and of heart sections after euthanization, we store measurements of the left ventricle pressure that syncs with images of the living heart. We look at the data we acquired and use our home-grown custom software to track particles on the surface of the heart to see how far and how fast those particles are moving.

The paper lab notebook basically performs the function of being an index into the actual datasets and it should record all the information the PI specifies. The content of a lab book relates to a particular experiment and is used by all staff working on that experiment. There could be on average 5-6 people using the notebook. We also have a paper surgical log that is kept with the animal and whatever project staff write down in that surgical log should be transferred into the lab notebook – so the data have to be in 2 places. They have to be kept with the animal in case there is a problem with the animal, but the PI also needs the data in the lab notebook to be able to write papers. Lab notebooks are just in the PI's office or lab with no backup.

1. Data ownership, privacy and ethical issues

a) What are the ethical or privacy issues in this study and how are they being addressed? What are the implications do these issues have for re-use?
b) What are the issues related to the home-grown custom software used for analysis in terms of potential future re-use?
c) Who will own any copyright or intellectual property rights to the lab notebooks, data sets or custom software code?
d) How might the data sets or custom software be licensed or fees charged if rights exist?
e) If re-use requires sharing of lab notebooks, how might this be managed? What would make re-use of lab notebooks easier for re-use?

Excerpt of Research Data Management Case C for use in Module 5

An MD applied for grant funding to do a qualitative study focusing on how to improve physician communication with African Americans (AA) and their relatives when their patients were receiving end-of-life care.

This qualitative study was conducted to expand knowledge about AA experiences and opinions about end-of-life care. Multiple-meeting focus groups were held to build trust and allow time for full participation. Following a review by a Community Advisory Board (CAB), protocols were approved by the University's Institutional Review Board. Participants were AA adults who had experienced at least one death of a significant other or family member

Data collection All participants gave informed consent. An open-ended interview script stimulated discussion about (1) positive and negative experiences of participants related to end-of-life care in the hospital or at home, (2) preferences for treatment by health care providers, (3) communication issues, and (4) end-of-life decision making pertaining to living wills and advance directives. An AA member of the project staff moderated the focus groups.

Each session was audio-taped taped. Unlabeled tapes were mailed to a transcriptionist in their plastic cases which were labeled. During the mailing process the package was damaged and the plastic tape cases broke and were no longer associated with the tapes for which the cases had been labeled. The tapes, however, were not damaged. The transcriptionist transcribed the tapes and the transcripts were sent back to the project team for identification of which focus group and which session should be used to identify each transcript. Focus Group Participants' comments were identified on the transcript by either Miss, Mrs. or Mr. plus the first initial of their first name. The transcripts were also reviewed for accuracy by the project team.

Data analysis Within-focus group multiple meeting set thematic analyses were performed, as well as cross-focus group set analyses to develop themes/recommendations for how end-of-life care communications might be conducted to improve the process for all concerned.

The tapes were eventually destroyed and the transcripts and other files generated during the analysis remained with the analyst who was not part of the project team and was affiliated with another medical school. The analyst was very involved with the drafting of the publication. Excerpts from the transcripts were later re-used as examples for a qualitative analysis class taught by the analyst at her medical school; however, for the reuse, all participant IDs were changed to P1, P2, etc.

1. Data protection/privacy

    a. What sorts of privacy conditions might a funder require for the data collected in this study?
    b. What ethical or privacy issues does this study present related to sharing or reusing data? How can they be resolved?
    c. What mechanisms were used to identify individual patients and maintain privacy? Would these need to be changed to preserve confidentiality during re-use?
    d. What issues might arise related to copyright or intellectual property rights to the data from each source or outcome data?
    e. How might the dataset be licensed or access fees charged to re-users if rights exist?

Excerpt of Research Data Management Case D for use with Module 5

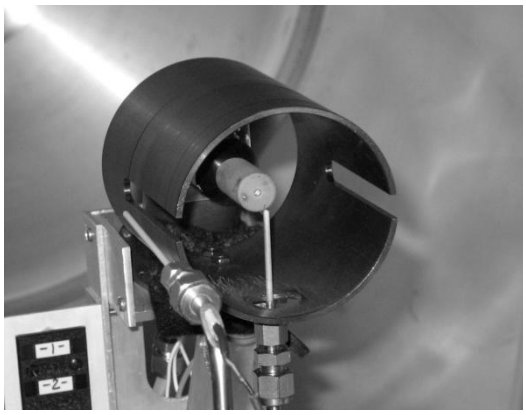


Fig. 1 50" x 72" vacuum chamber used for cathode research (WPI Higgins Laboratory)

Fig.2. Hollow cathode and surrounding anode installed in vacuum chamber. Also visible are Langmuir and emissive probes (small white rods).

Scenario:

A faculty researcher in Aerospace Engineering studies electric propulsion for spacecraft control, in other words a type of rocket engine that uses electricity to ionize and accelerate a gas to produce thrust. He is concerned about data security in his lab and is looking for a standard protocol that could be used by all and would comply with any security requirements imposed by his research sponsors.

The goal of one of his current projects is to study and characterize a component of an electric thruster being used by NASA, the Air Force, and private companies to control satellites in orbit. This work enables researchers to build more robust thrusters that will have a longer service life than current models, thus enabling longer and more ambitious space missions. The students in the lab have been experimenting on a particular engine component, called a "hollow cathode," to characterize the plasma it generates. This data will help researchers understand where energetic ions are produced that erode surfaces and limit the cathode lifetime. They perform experiments using Langmuir and emissive probes to collect data from two different cathodes tested in a vacuum chamber.

The two cathodes used in the lab have different restrictions on their use. The first is from a private company, was developed with Air Force funding, and is the same model as a unit which has been used operationally in orbit. Because it is identical to flight hardware, work with this cathode must comply with International Traffic in Arms Regulations (ITAR) requiring that no foreign nationals have access to any aspect of the research. ITAR establishes strict controls on the use and dissemination of information related to defense articles. Some equipment and lab notebooks are kept locked up in order to comply. The second cathode is from NASA, is a laboratory use model (i.e. not "flight hardware"), and is not subject to the same ITAR restrictions.

Raw data is generated from the cathodes during experiments and is downloaded onto a laptop. For useful data, the students produce code using MATLAB and create appropriate graphs and charts displaying the data points. The MATLAB code is the intellectual property of the student

and faculty advisor who create it. If code is written primarily by a student, then the advisor will usually try to contact the student (who may have graduated) for permission before distributing it to another research group. Permission is rarely granted for sharing this code outside the research group. When students graduate they are required to provide the researcher with a CD containing the MATLAB code used for any analysis they completed in the lab.

Laboratory notebooks which include any information on projects which are ITAR restricted are kept in a secure location, either in the researcher's office or in a locked storage cabinet located in a lab. The archive CDs are stored in the researcher's office and there is no backup plan for either the CDs or the notebooks.

---

1. Data Privacy/Restrictions

   a. What specific actions have been taken by the researcher in order to comply with ITAR restrictions?

   b. What concerns could arise related to privacy of research being conducted for a private company? How could these be addressed?

   c. Describe any intellectual property and ethical concerns that could arise related to the MATLAB code created by the researcher's students.

Assessment Questions for Research Data Management Case Excerpts for Module 5

A: Outcomes from Orthopedic Implant Surgery

Select all response options that apply

1. The owner of the data in this case is:
    a. The original research resident
    b. The last research resident
    c. Dr. X
    d. The institution where Dr. X is employed
    e. All of the above
    f. None of the above

2. Ethical issues related to re-use of data in this case might include
    a. Providing a copy of the proprietary software package used to monitor patients walking rate to a researcher who wants to do a follow-on study
    b. One of the residents copying the data for later follow-on research at another institution
    c. Using the data which has been transferred to a public repository as part of a later study at the same institution and publishing a paper citing a publication related to the earlier study
    d. All of the above
    e. None of the above

3. Legal issues related to this case include:
    a. Providing a copy of the proprietary software package used to monitor patients walking rate to a researcher who wants to do a follow-on study
    b. Copying the analytic software for the accelerometer data to other computers outside the lab
    c. Providing the survey questions for re-use outside the institution
    d. All of the above
    e. None of the above

4. Privacy issues related to this case might include:
    a. Giving surgeon notes from physical exams to another researcher for reuse
    b. Sharing structured database of physical exam measures for reuse
    c. Providing patient survey data for re-use
    d. Providing the data generated by the accelerometer for re-use

e. Use of all of the data for a later study on pain and disability following any kind of orthopedic surgery
f. All of the above
g. None of the above

Answers for Case A:

1. d: The institution where Dr. X is employed owns the data and Dr. X is the steward of the data

2. d: All of the above: a) Proprietary software requires a separate license for each PC on which it is used . b) A resident could not copy the data without permission from the home institution and Dr. X. c) If the data has been published in a public repository, the data should be cited as well as the prior publication

3. a&b: a) Proprietary software requires a separate license for each PC on which it is used. Survey questions, if they are part of a copyright protected survey tool could probably not be re-used without compensating the originator; however, if it is a survey that has been developed in house or has never been given copyright protection, it could be reused with permission of the developers.

4. f: All of the above:
   a-d) Use of any of the data sets would require permission from the home institution unless they are published in a public repository. e) Additionally, the informed consent originally signed by the patients would have to have asked them to agree to use of the data for purposes other than the original study. If the data were to be reused, patient data needs to be stripped of personal identifiers.

Case B: Regeneration of Functional Heart Tissue in Rats

Select all response options that apply

1. The owner of the data in this case is:
    a. The principal investigator
    b. The project director
    c. The senior project staffer
    d. The institution where the principal investigator is employed
    e. All of the above
    f. None of the above

2. Ethical issues related to re-use of data in this case might include
    a. Providing a copy of the custom software used to track particles on the surface of the heart to a researcher at another institution who wants to do a follow-on study
    b. One of the project staff members copying the data for subsequent related research at another institution
    c. Using the data as part of a later study at the same institution and publishing a paper citing a publication related to the earlier study
    d. All of the above
    e. None of the above

3. Legal issues related to this case include:
    a. Providing a copy of the custom software used to track particles on the surface of the heart to a researcher at another institution who wants to do a follow-on study
    b. Copying the custom tracking software to other computers outside the lab within the home institution
    c. Providing the images taken of the rat heart which have been placed in a public repository for re-use outside the institution and publishing a paper citing a publication related to the earlier study
    d. All of the above
    e. None of the above

4. Privacy issues related to this case might include:
    a. Leaving the surgical log in the unlocked lab
    b. Sharing database of left ventricle pressure measures with synchronized images of the living heart for reuse
    c. Providing the images of the slides containing stained heart sections for re-use
    d. All of the above
    e. None of the above

Answers for Case B:

1. d) The institution where the principal investigator is employed owns the data and the PI is the steward of the data

2. d: a) custom software cannot be reused without permission and acknowledgement of the source unless it is available under an open source license. b) a resident could not copy the data without permission from the home institution and PI. c) If the data has been published in a public repository, the data should be cited as well as the prior publication

3. a&c: a) custom software cannot be reused without permission and acknowledgement of the source unless it is available under an open source license. b) since the custom software does not require a proprietary license, being homegrown, it can be used on other computers at the home institution with the developer's permission. c) If the data has been published in a public repository, the data should be cited as well as the prior publication

4. e: None of the above: None of these are privacy issues since human subjects are not involved as participants in the study; however, there may be security issues with option a) and legal issues with b) and c).

Case C: Improving End-of-Life Care for African Americans

Select all response options that apply:

1. The owner of the data in this case is:
    a. The MD who applied for the grant funding the study
    b. The African-American adults who participated in the focus groups
    c. The moderator of the focus groups who was part of the project team
    d. The institution where the MD principal investigator is employed
    e. The qualitative analyst who generated the findings for the study
    f. All of the above
    g. None of the above

2. Ethical issues related to re-use of data in this case might include
    a. Audio-recording the focus groups after informal consent from participants during their recruitment by telephone
    b. Publishing a paper detailing the results of study without having the paper reviewed by the focus group participants
    c. Having focus group participants sign an informed consent form after the focus group
    d. All of the above
    e. None of the above

3. Legal issues related to this case include:
    a. Providing to attendees at a professional conference presentation a pre-publication copy of a manuscript detailing results of the study submitted to a journal that is not open access
    b. Copying the proprietary qualitative data analysis software to a computer at the analysts' institution
    c. Providing the focus group moderator script for re-use outside the institution
    d. All of the above
    e. None of the above

4. Privacy issues related to this case might include:
    a. Providing a copy of the audiotapes used to record the focus groups to a researcher who wants to do a follow-on study
    b. Providing a copy of the data transcripts for later follow-on research at another institution
    c. Use of all of the data for a later study on comparing end-of-life care considerations across cultures (Anglo, Latino, and African American)
    d. Storing transcripts and analytic data structures with the analyst at another institution
    e. Use of de-identified segments of the transcripts as examples for a qualitative analysis class
    f. All of the above

g. None of the above

Answers for Case C:

1. d) The institution where the MD principal investigator is employed. The MD is the steward of the data

2. a&c) a written and signed informed consent form is required by most Institutional Review Boards and this document must be signed before recording participants so they know they will be recorded and agree to this. b) part of the informed consent generally states that study results may be published and that no participants will be identified by name, roles, or any other non-generic personal characteristics in any publication

3. a&b) a) Generally pre-publication copies of manuscripts are not to be circulated b) Proprietary software requires a separate license for each PC on which it is used

4. a,b&c) Unless the participant informed consent form explicitly asked for permission to reuse data in other studies, this should not be done. However, if it is de-identified (i.e., participants are not identifiable), project staff may use the data for teaching purposes. d) Storing the data and intermediate results with a staff member at another institution without keeping a backup copy at the home institution could pose a loss of data risk, but all project team members have been generally been certified to do research with human subjects and adhere to privacy constraints

Case D: Characterizing a Component of a Rocket Engine used to Control Satellites in Orbit

Select all response options that apply:

1. The researcher has taken the following actions to comply with ITAR restrictions:
   a. He limits use of ITAR restricted equipment to PhD students only.
   b. He does not allow any students who are foreign nationals to have access to ITAR restricted equipment, notebooks or research materials.
   c. He secures the ITAR restricted equipment but is able to share the lab notebooks and any other research documentation.
   d. All of the above
   e. None of the above

2. Intellectual property issues related to this case might include:
   a. Restrictions on publishing research that has been completed for a private company.
   b. Moving raw data from a laptop to an archival CD.
   c. Inability to share MATLAB code that has been created by a student.
   d. All of the above
   e. None of the above

3. Which of the following data is owned or can be reused by the researcher without requesting permission:
   a. The lab notebook
   b. Raw data downloaded from the cathodes onto a student's laptop
   c. Archival CDs of MATLAB code
   d. All of the above
   e. None of the above

4. The researcher could do which of the following to clarify intellectual property issues related to the MATLAB code:
   a. Continue with current practice. No clarification is needed.
   b. Create two CDs including the code. One would be owned by the student and one by the researcher.
   c. The researcher can request that students sign a form granting reuse right to the researcher prior to graduation.
   d. All of the above
   e. None of the above

Answers for Case D:

1. b) Foreign nationals may not have access to any part of an ITAR restricted research project.

2. a&c) Information that is owned by a student or a private company may not be shared without permission of the owner.

3. a&b) The lab notebook is the intellectual property of the researcher and raw data is not governed by copyright laws in the United States. The MATLAB code is the intellectual property of the student so permission must be sought in order to reuse or adapt it.

4. c) Option b would not have any impact on who owns the intellectual property contained on the CD. If the student has given consent to the research to use the MATLAB code and adapt it as needed, there would be no need to track down the student later, therefore simplifying the process.

Appendix A: Roster of Steering and Education Committees and Consultants

The *Frameworks for a Data Management Curriculum* packet has been developed through the collaborative efforts of the following project committees and consultants:

Steering Committee

- Elaine Martin, D.A., MSLS, Co-Chair
- Tracey Leger-Hornby, PhD, Co-chair
- Sia Najafi, MS, Director of Research Computing, Worcester Polytechnic Institute
- Mary Piorun, MSLS, MBA, Associate Director, Lamar Soutter Library, University of Massachusetts Medical School
- Donna Kafel, MLIS, Project coordinator, University of Massachusetts Medical School

Education Committee Members

UMass Medical School	WPI
Patricia Franklin, MD, MBA, MPH	Christine Drew, MLS
Donna Kafel, MLS	Glenn Gaudette, Ph.D.
David Lapointe, Ph.D.	Laura Hanlan, MLIS
Myrna Morales, MAT, MSLIS	Erica Stults, MS, Ph.D. candidate
Lisa Palmer, MSLS	John Sullivan, D.E.

Project Consultants

- Curriculum Design: Paul Colombo, MS
- Evaluation Expert: Nancy LaPelle, Ph.D.
- Instructional Design: Heather McMorrow, MA