Syllabus for Data, Data Practices, and Data Curation Part 2; Spring 2012

Christine L Borgman, University of California, Los Angeles

Available at: https://works.bepress.com/borgman/261/
Course Description, Parts I and II

In today’s technology-intensive research environments, petabytes of data may be produced in a matter of days, weeks, or months. Those data may be lost in a similar amount of time if they are not captured, curated, and marked up in ways that allow for discovery and reuse by others. Datasets large and small can be very useful not only to researchers, but also to students, to the general public, and to policy makers. Among the classes of data of broad general interest are scientific records of the climate, the skies and galaxies, plant and animal species, social and economic observations, and cultural and historical records. Research policy by governments and funding agencies encourages – and increasingly requires – that investigators make plans for data management, curation, and dissemination.

The National Science Foundation announced a new requirement in 2010 for all grant proposals: they now must include data management plans. This requirement is causing a mad scramble for compliance by researchers, universities, librarians, and archivists. The Association for Research Libraries already has issued several reports on the library’s role in data management, adding to a plethora of policy reports in this area. The Institute for Museum and Library Services is funding curriculum development for data librarians. Data management is clearly a growth area for information studies graduates.

These two courses (winter and spring) will survey the rich landscape of data practices and services, including data as evidence and their role in research; data-intensive research methods; social studies of data practices; national and international data policy (e.g., intellectual property, release policies, open access, economics); comparisons between disciplines; management of data by research teams, data centers, libraries, and archives; technical standards for data and metadata; and data curation. Part I (winter) lays the foundation for data practices and services across the disciplines. Part II (spring) builds upon this background to provide practical experience in data curation. One large project will be undertaken across the two terms plus several smaller assignments. The courses will be graded separately. Part I is a pre-requisite for Part II. However, by taking Part I, you are not obligated to take Part II.

These courses will be a mix of readings, discussion, practicum, field trips, and guest lectures. Invited speakers for this term include local experts and distinguished guests from the National Academy of Sciences and the Library of Congress by videoconference.
Librarians, archivists, and other information professionals bring essential skills to the realm of research data. Information activities related to data include developing metadata, standards, and systems of classification, establishing archival plans for data selection, migrating data to new platforms and standards, creating finding aids for multiple user communities, and developing databases and technologies to support data creation, preservation, discovery, and reuse. Funding agencies and faculty are looking to libraries for leadership for the management, curation, hosting, and dissemination of research data. Data librarianship is a growth area in academic and special libraries, and will be an increasingly important set of skills for librarians and archivists in all sectors.

This is an introductory graduate course, suitable for masters and doctoral students in information studies and in data-intensive research fields. The course is open to practicing librarians and archivists through concurrent enrollment, with instructor’s permission.

The two-part sequence of courses in Data, Data Practices, and Data Curation has been developed with the substantial contributions of UCLA doctoral students Jillian Wallis and Laura Wynholds and guidance from students enrolled in prior offerings of the course in 2010 and 2011. Thanks also are due to instructors of similar courses at other universities who shared their syllabi and course materials, especially Margaret Hedstrom and Ann Zimmerman of the University of Michigan, Carole Palmer and Melissa Cragin at the University of Illinois, and Carolyn Hank at the University of North Carolina.

**Course Objectives**

1. Students will learn to distinguish between the many forms of data, how data vary by scholarly discipline, and how they are used throughout the scholarly life cycle.
2. Students will learn some professional criteria for selecting and appraising data.
3. Students will learn to distinguish among different types of data collections, repositories, and services.
4. Students will learn the roles that data play in research collaborations.
5. Students will gain a basic knowledge of data curation practices in the library and archive fields.
6. Students will learn basic principles of public policies for data.
Course Materials

All readings and other course materials will be posted on the Moodle site for this course. Enrolled students have access to the site at http://www.ccle.ucla.edu.

Office Hours

Most Tuesdays, 1:30-3pm (link posted on CCLE), other times by appointment, and always by email.

Grading

Short paper assignment 25%
Term project 50%
Class participation and analysis of readings 25%

Details of the assignments are provided on separate documents.

Students are expected to complete all assigned the readings prior to each week’s class sessions and come prepared to discuss them. Your preparation and contributions to the discussion are the basis for 25% of your grade. Written assignments are due at the beginning of the class session, on paper, and are to be submitted electronically to the CCLE / Moodle site. Assignments will be marked down 2 points for each day late. No assignments will be accepted after midnight on TUESDAY, JUNE 12.
Overview:

Course Description, Parts I and II .......................................................... 1
Course Objectives ............................................................................. 2
Course Materials .............................................................................. 3
Office Hours ..................................................................................... 3
Grading ............................................................................................... 3
Overview: ........................................................................................... 4
Topics, Readings, and Guest speakers .................................................. 5
  Week 1: Course Introduction, April 5 ................................................. 5
  Week 2: Selection and appraisal, April 12 ......................................... 7
  Week 3: Repositories, April 19 ......................................................... 8
  Week 4: Collection management, April 26 .................................... 10
  Week 5: Statistics, computation, and trust in data, May 3 ............... 11
  Week 6: Economics of data curation, May 10 ................................. 14
  Week 7: Formal representations of data, May 17 ............................. 15
  Week 8: Identifiers and data citation, May 24 ................................. 18
  Week 9: Provenance, May 31 .......................................................... 20
  Week 10: Project presentations, June 7 ........................................... 22
Topics, Readings, and Guest speakers

Week 1: Course Introduction, April 5

This week’s readings provide an overview of data curation issues. In addition to a discussion of these readings, we will provide an overview of topics, assignments, and course goals with the intent of bringing everyone to a common base, as some students took Part I in Winter 2011 and some in Winter 2012. Note that one of the “readings” is a video of an hour or so.

Please also read and contribute to the California Digital Library Data Pub Blog, which may be found at: http://datapub.cdlib.org. We will also be discussing the Dataverse Network, which may be found at: http://www.thedata.org.

We will form project groups for the term today, so please review the assignments and be prepared to discuss your preferences.

Readings:
Recommended:


Week 2: Selection and appraisal, April 12

How do we determine what data are worth keeping? What are the criteria? To what extent are these scholarly, disciplinary, policy, or professional questions? These are core questions in both libraries and archives, thus we draw readings from each tradition to compare approaches. See also the chart of comparisons we developed last year (posted on CCLE).

Guest discussant: Laura Wynholds, UCLA PhD student on Data Practices Research Team

Readings:


Recommended:


Week 3: Repositories, April 19

Data collections and repositories take many forms and reside in many places. Libbie Stephenson will provide an introduction to the technical aspects of data ingest and professional aspects of repository management. In addition to the readings below, please review the policy documents posted on the CCLE site for this week, provided by Ms. Stephenson.

**Speaker:** Libbie Stephenson, Director, UCLA Social Science Data Archive  
**Guest:** Dr. Marina Jirotka, Associate Director, Oxford e-Research Centre

**Readings:**


**Recommended:**


Science special issue on Data, Feb 11, 2011- pick some articles. Retrieved from http://www.sciencemag.org/content/331/6018.toc


We continue the repositories discussion with an assessment of collections policies. Not all data can be accepted into collections, and once accepted, they must be kept secure and kept accessible. These are complex problems with competing perspectives, as reflected in the choice of readings.

**Readings:**


**Recommended:**


“Trust” pervades the life cycles of data. Researchers learn how to trust their own data and the data of others. They have scholarly mechanisms to assess whether data or findings are trustworthy. They make choices about whether to trust repositories as places to put their data or from which to retrieve. Similarly, archivists and data repository managers have established practices for how to ensure trust in repositories and in the data they hold. Some aspects of trust involve tacit knowledge about scholarship, some involve technologies, and some involve statistical analysis. We will explore these various notions of trust with our guest speaker, Prof. Victoria Stodden of the Dept. of Statistics at Columbia University, and visiting scholar at the UC Berkeley I-School.

**Speaker:** Prof. Victoria Stodden, Columbia University, Dept. of Statistics

**Readings:**


**Recommended:**


Week 6: Economics of data curation, May 10

Data are expensive to create, manage, and curate. Losing access to data also can be expensive. Economic models for managing digital data are proliferating, each with its own set of assumptions. We will have a panel discussion with several distinguished speakers involved in economic aspects of data curation management and policy. This session will be held in the large conference room (11360) of the Young Research Library and open to all. Attendees are encouraged to read the background materials prior to the session.

Format of the session:
9-9:30 Overview of issues and readings: Prof. Borgman
9:30-10:50: 15-min presentations by each speaker, with questions to each
9:30-9:50: Tony Hey on economics of cyberinfrastructure to support data-intensive research
9:50-10:10: Stephen Abrams on economic/business model for data curation of the California Digital Library
10:10-10:30: Todd Grappone on UCLA economic/business models for data curation
10:30-10:50: Jessie Hey on university repository economics for data break
11:10-12: general panel discussion on economics of data curation

Speakers:
Stephen Abrams, California Digital Library, Merritt repository
http://www.cdlib.org/cdlibinfo/2009/11/19/3052/
Dr. Tony Hey, Corporate Vice President of Microsoft Research; formerly Director of eScience Initiative, UK: http://www.microsoft.com/presspass/exec/tonhey/
Dr. Jessie Hey, formerly ePrints repository, University of Southampton, UK

Readings:


Recommended:


JISC. (2012). Text mining promises huge economic and research benefit, but copyright law and other barriers are limiting its use, says JISC report. JISC. Retrieved from http://www.jisc.ac.uk/news/stories/2012/03/textmining.aspx


Week 7: Formal representations of data, May 17

While metadata and other formal representations for data are important aspects of curation, they are applied inconsistently across types of data and communities of practice. We will discuss representations, ontologies, and standards in the ideal and in the practical world.

Speakers: Jillian Wallis & Laura Wynholds, UCLA

Readings:


Week 8: Identifiers and data citation, May 24

Data need to be identified in ways that are unique and that are persistent. Data citation and attribution are becoming critical issues in data curation and access. The CODATA report from the National Academies is due to be released by the date of the class; it will replace the slide presentations listed here.

Readings:


ORCID or how to build a unique identifier for scientists in 10 easy steps - Gobbledygook Blog.
Nature Publishing Group. http://blogs.nature.com/mfenner/2010/01/03/orcid-or-how-to-build-a-unique-identifier-for-scientists-in-10-easy-steps. See also Open Research and Contributor ID http://www.orcid.org/

Recommended:


Week 9: Provenance, May 31

While metadata and other formal representations for data are important aspects of curation, they are applied inconsistently across types of data and communities of practice. We will discuss representations, ontologies, and standards in the ideal and in the practical world.

Speaker: Yolanda Gil, Information Sciences Institute, USC

Readings:
Recommended:


Week 10: Project presentations, June 7

See project assignment for details. We will devote the last class session to a public presentation of student projects and to a general discussion of the data curation issues identified in each project.