Data, Data Practices, and Data Curation

Part I, Winter, 2012, UCLA Information Studies
Tuesdays 1:30-4:50, IS Room 121
Christine L. Borgman, Professor & Presidential Chair
215 GSE&IS bldg, 310-825-6164; borgman@gseis.ucla.edu

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

Wednesdays, 10am-noon (link posted on CCLE), other times by appointment, and 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 Wednesday, March 21, without advance permission from the instructor.
Overview:

NOTE: This document is current as of early January. The order of topics may change depending upon the availability of guest speakers. In case of conflicts, the CCLE site takes precedence. An updated version of the syllabus will be posted when all speakers are finalized.
Week 1: Overview of Data, Data Practices, and Data Curation, January 10

We will devote the first week of class to an overview of the concept of data and its manifestations across scholarly disciplines and in public parlance. We will also begin to form project groups for the term.

Readings are to be completed in advance of each class session. Please come to class prepared to discuss the material and its relationship to larger issues in the course and the curriculum. Prepare some talking points as part of your reading and studying.

**Required readings:**

http://rsta.royalsocietypublishing.org/content/368/1926/4023.full

http://www.sciencemag.org/content/323/5919/1297.full.pdf

**Recommended readings:**


Week 2: What are data? January 17

“Data” is a far more ambiguous concept than is immediately apparent. Decisions about what data are to be managed, shared, and curated depend heavily on how the concept is defined. We will devote today to exploring some of the many definitions and facets of “data.”

Assignment: Bring in a sample today of something that you consider to be data. We will discuss them in class.

Readings:
Chapter 1: Importance and use of scientific and technical databases, p 14-29.

Recommended readings:
Week 3: Roles of data in research, January 24

The best way to understand research data as to observe how data are created, used, managed, analyzed, and published. The readings address an array of disciplines and examples of data use. Our two guest speakers will provide contrasting examples of data context representing distinctly different disciplines, collection methods, and publication venues.

Speakers:

Katie Mika, UCLA PhD candidate in Civil and Environmental Engineering. *Widespread and high levels of fecal indicator bacteria do not correlate with hotspots of human fecal source contamination in the Santa Monica Canyon Channel*

Kimberly Garmoe, Recent UCLA PhD graduate in History. *Frankendata*

The class session will be structured as follows:
1:30-2:30 discussion of this week’s readings
2:30-2:45 break
2:45-3:30 Katie Mika (presentation and discussion)
3:30-4:15 Kimberly Garmoe (presentation and discussion)
4:15-4:45 general discussion and comparisons
4:45-4:50 next week’s readings, etc

Readings:

**Recommended:**

Week 4: Data release and sharing, January 31

Librarians, archivists, researchers, and policy makers are concerned with how to capture, manage, and curate data. If data are to be available for curation and reuse, they must be released by the researchers who collect, produce, or hold them. Notions of data sharing varies widely by individual, group, and research specialty, as well as by funding agency.

Readings:

Read at least two of these:

Recommended:


Week 5: Public policy for research data, February 7

The ability to deposit, discover, share, retrieve, reuse, and curate data all depend upon public policies about rights and responsibilities. These policies have legal and economic aspects and vary widely around the world, although many international agreements are in place.

Speaker (by video): Paul Uhlir, Board on Research Data and Information, National Academies of Science

NOTE: Today class meets in the new conference room in Young Research Library, 11360.

Readings:


Recommended:


Week 6: The role of data in replicating and reproducing research, February 14

The need to replicate or reproduce research is a common justification for data release, sharing, and curation. However, notions of replication and reproducibility are as problematic as are definitions of “data.”

Readings:


Read at least two of these articles:


**Recommended readings:**


Data may be deposited in archives or repositories of various kinds, organized by discipline, content, institution, or other criteria. Data repositories serve similar roles to other kinds of repositories, providing permanent access and curation.

**Speaker:** Elizabeth Stephenson, Director, [Institute for Social Research Data Archive](http://www.isrdataarchive.org), UCLA

**Readings:**


**Recommended:**


Week 8: Collaboration and life cycles, February 28

Throughout the stages of a research project, data may be identified, captured, analyzed, modeled, modified, or discarded. Maintaining records of these stages and the relationships between them can be critical for data curation. Research teams handling data may vary in size from a few people to many hundreds. Data are also the “glue” that holds these teams together.

Readings:

Recommended:
Week 9: The Role of Libraries & Archives in Data Management, March 7

Only recently have research libraries begun to consider their roles and responsibilities in the curation and management of research data. We will survey current practice, policy, and proposed directions for data services provided by libraries and archives. Peter Young, Chief of the Asian Division of the Library of Congress, is leading LC’s efforts in eScience and data curation, including the Twitter archive. Mr. Young and some of the LC staff will join us by videolink.

Invited Speaker (by video): Peter Young, Library of Congress

Class will meet in YRL 1160.

Readings:
Week 10: No Class session, March 13
You will have this week to work on your final project and your presentations for next week.

Week 11 (exam week): Student presentations, March 20
See project assignment for details. We will devote the last class session to a public presentation of student projects, and a general discussion of project findings.

Wednesday, March 21: Final projects due, 5pm, to instructor’s mailbox and by PDF to CCLE.