Data, Data Practices, and Data Curation

Part II, SPRING 2014, UCLA Information Studies
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 was 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, 3:00-4:30pm (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 site. Assignments will be marked down 2 points for each day late. No assignments will be accepted after midnight on Weds, JUNE 7.
### 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, April 2: Selection and appraisal .......................................................... 6
Week 2: April 9: Data archives and repositories .............................................. 7
Week 3: April 16: Economics of data: preservation, access, and sustainability ... 8
Week 4: April 23: Managing the Data Deluge: NISO seminar ............................ 10
Week 5: April 30: Provenance in data: WWW and archival approaches, Paul Groth, University of Amsterdam, by video ................................................................. 11
Week 6: May 7: U.S. and international data policy: Paul Uhlir, National Academies of Science, by video ................................................................. 13
Week 7: May 14: Intellectual property in data: Maureen Whalen, Getty, in person and Peter Hirtle, Berkman Center, Harvard, by video ........................................................................ 15
Week 8: May 19: Research Libraries Respond to Data Policy: MacKenzie Smith, University Librarian, UC-Davis (Special Session) ................................................................. 17
Week 9: May 28: Research Managerial issues in data: research libraries and archives, invite representatives from UCLA libraries and CDL ........................................................................ 18
Week 10: Project presentations, June 7 .......................................................... 20
Week 11 (Exam Week): June 11: Projects due ................................................. 20
Topics, Readings, and Guest speakers

Week 1, April 2: Selection and appraisal across domains

Week 2: April 9: Data archives and repositories

Week 3: April 16: Economics of data: preservation, access, and sustainability

Week 4: April 23: Managing the Data Deluge: NISO seminar
http://www.niso.org/news/events/2014/virtual/data_deluge/

Week 5: April 30: Provenance in data: WWW and archival approaches, Paul Groth, University of Amsterdam, by video

Week 6: May 7: U.S. and international data policy: Paul Uhlir, National Academies of Science, by video

Week 7: May 14: Intellectual property in data: Maureen Whalen, Getty, in person and Peter Hirtle, Berkman Center, Harvard, by video

Week 8: May 19: Research Libraries Respond to Data Policy: MacKenzie Smith, University Librarian, UC-Davis (Special Session)

Week 9, May 28: Technologies of persistence and identification

Week 10: June 7: Project presentations

Week 11 (Exam Week): June 11: Projects due
Week 1, April 2: Selection and appraisal

How do we determine what data are worth keeping? What are the criteria? To what extent are these scholarly, disciplinary, policy, or professional questions? Who decides? What role do scholars play in determining what is kept and what are professional roles?

Readings

(Borgman, 2015), Chapter 10, What to Keep and Why
(Faniel & Jacobsen, 2010)
(Gutmann et al., 2009)
(Harvey, 2008)
(“Strategic Directions: Appraisal Policy,” 2007)
(Goodman et al., 2014)
(“Australian Code for the Responsible Conduct of Research,” 2007)
(Pfeiffenberger & Carlson, 2011)

References

Week 2: April 9: Data archives and repositories

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.

**Speaker: Libbie Stephenson, Director, UCLA Social Science Data Archive**

**Readings (and sites to visit)**

Only a couple of these are narrative readings (Greenberg, White, Carrier, & Scherle, 2009; Kimpton & Morris, 2013; Peer, Green, & Stephenson, 2014). The COAR report (“COAR: The Current State of Open Access Repository Interoperability,” 2012) should be skimmed, and we’ll return to it later when we discuss technologies for access to data. The rest are sites that you should visit and browse to get a sense of the array of data repositories and issues such as standards and certifications.

**Open Data**
Open Data Foundation [http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers](http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers)

**Standards**

**Metadata**

**Repositories – Management issues**
(Kimpton & Morris, 2013; Peer et al., 2014)

**Repositories - offering services**
UK Data Service [http://ukdataservice.ac.uk/](http://ukdataservice.ac.uk/)
Week 3: April 16: Economics of data: preservation, access, and sustainability

In managing and curating research data, the economic issues are unavoidable. What are the economic models for sustaining access to data? How can projects be costed and budgeted? How do economic considerations influence choices of what to keep, for how long, and why? We will consider several approaches to the economics of research data.

Readings

(Beagrie & Houghton, 2014)
(Berman et al., 2010)
(David, 2009)
References


Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of “open science.” BioSocieties, 7(2), 140–168. doi:10.1057/biosoc.2012.8


Recommended Readings


Week 4: April 23: Managing the Data Deluge: NISO seminar

In class attendance of the NISO Virtual Conference: Dealing with the Data Deluge: Successful Techniques for Scientific Data Management.
http://www.niso.org/news/events/2014/virtual/data_deluge/

The class today will be held in YRL 11348, the presentation room, from 8am to 1pm. Students are encouraged to come early to hear the keynote speaker (8-9am) and stay for the closing session if they can. Because Prof. Borgman is among the presenters (11:15-11:45am), the fees are waived for participants.

Readings for this week are by the presenters or on their topics, so that you will have a full framing of the issues being addressed by these leaders in the field.

About the Virtual Conference (NISO text)

With the expansion of digital data collection and the increased expectations of data sharing, researchers are turning to their libraries or institutional repositories as a place to store and preserve that data. Many institutions have created such data management services and see the data curation role as a growing and important element of their service portfolio. While some of the experience in managing other types of digital resources is transferrable, the management of large-scale scientific data has many special requirements and challenges. From metadata collection and cataloging data sources, to identification, discovery, and preservation, best practices and standards are still in their infancy.

This Virtual Conference will explore in greater depth than traditional webinars some of the practical lessons from those who have implemented data management and developed best practices, as well as provide some insight into the evolving issues the community faces. It will include discussions related to certification of trusted repositories, provenance and identification issues around data, data citation, preservation, and the work of several repository networks to advance distribution of scientific information.

Readings

(CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013)
(Crosas, 2011)
(Datacitation Synthesis Group, 2014)
(DataCite, 2013)
(“The Dataverse Network: Presentations,” 2012)
(The Purdue University Research Repository (PURR), 2013)
(Holdren, 2013a, 2013b)

References

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, 1–75. Retrieved from https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article


*The Purdue University Research Repository (PURR)*. (2013). Retrieved from http://www.youtube.com/watch?v=Yw0IJj7FqA8&feature=youtube_gdata_player

---

**Week 5: April 30: Provenance in data: WWW and archival approaches, Paul Groth, University of Amsterdam, by video**

Archival and Internet approaches, Paul Groth, University of Amsterdam, discussant, 9am-10:30am by video. Groth will discuss which of the many competing definitions of provenance is most useful for data, what aspects of provenance are most applicable to data management, to curation, to discovery, and to curation, what are the points of intersection between archival, historical, and technical concepts of provenance, especially with respect to data, and what are the practical considerations in documenting provenance for data stewardship.

**Readings**

(Barbier, Feng, Gundecha, & Liu, 2013)
(Bunn, 2014)
(Cheny, Gil, Groth, & Miles, 2011)
(Garber, 2012)
(Goodman et al., 2014)
(Groth & Moreau, 2013)
(Meng, 2011)
(Niu, 2013)

**References**


**Recommended Readings**


Week 6: May 7: U.S. and international data policy: Paul Uhlir, National Academies of Science, invited

Paul Uhlir will discuss the most pressing policy questions for research data in the U.S., how U.S. research policies for data differ from those of other countries and regions, such as UK, EU, and Australia, and his recent book on governance policy for microbial data.

Readings

(Anderson, 2013)
(Boulton, 2012)
(Burwell, VanRoekel, Park, & Mancini, 2013)
(“Data Archiving Policy,” 2012)
(“Joint Data Archiving Policy,” 2012)
(Duke & Porter, 2013)
(Holdren, 2013a, 2013b)
(Lide & Wood, 2012)
(Uhlir, 2007)

References


Recommended Readings


CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, 1–75. Retrieved from https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article


Week 7: May 14: Intellectual property in data: Maureen Whalen, Getty, and Peter Hirtle, Berkman Center, Harvard, by video

Policies and practices for open data tend to say little about the problems of intellectual property rights. The OECD report (read earlier in the term) is an exception. Yet rights in data are unclear and often contested. Investigators may not know what rights they have in data, and they may be obtaining data from other sources for which they have little control over rights, formats, reuse, deposit, etc. Complicating matters further are the license stacking problems, wherein each dataset has particular rights associated with it, making data difficult to combine.

Readings

(Ball, 2012)
(“Defining ‘non-consumptive’ use,” 2014)
(“LIBER response to Elsevier’s text and data mining policy | LIBER,” 2014)
(Hirtle, 2011)
(Scotchmer, 2003)
(Shillum, 2014)

References


Recommended Readings

Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of “open science.” BioSocieties, 7(2), 140–168. doi:10.1057/biosoc.2012.8


---

**Week 8: May 19: Research Libraries Respond to Data Policy: MacKenzie Smith, University Librarian, UC-Davis**

Special session, Monday, May 19, 12:30-3:30pm, in GSEIS room 121 with respondents Dr. Vessela Ensberg, Data Curation Analyst, UCLA Libraries, and Edson Smith, data curation strategist and technical lead on data curation, UCLA Libraries. Will discuss how universities are adapting to an environment of open access and open data mandates, how library associations such as ARL are addressing data management, what approaches UC campuses are taking, what role libraries play in research policy compliance, and similar topics.

**Readings**

(Crosas, 2013)
(IAU Working Group Libraries, 2013)
(Kenney, 2014)
(Smith, 2010)
(Soehner, Steeves, & Ward, 2010)

**References**


**Recommended Readings**

Week 9: May 28: Technologies of persistence and identification

The technologies associated with data curation and management vary widely by domain, method, and many other factors. We will focus on generic tools and principles for technology as an entry point for developing a broader background. These readings and the video address the following topics:

- Persistent identification
- Unique identification of digital objects
- Digital Object Identifiers
- CrossRef
- Open Archives Initiative protocols OAI-PMH and OAI-ORE
- Research Objects and provenance
- Linked open data for data
- Namespaces, URLs, and versions of record

Readings

(Bechhofer et al., 2013)
(Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010)
References


Recommended Readings


**Week 10: June 7: Project presentations**

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.

**Week 11 (Exam Week): June 11: Projects due**

Final projects are due by 5 p.m. on Wednesday, June 11.