Data Curation and Policy

Winter, 2015, UCLA Information Studies 262B,
Weds, 9am-12:20pm, IS Room 245,
Christine L. Borgman, Professor & Presidential Chair
235 GSE&IS bldg, 310-825-6164; Christine.Borgman@ucla.edu

Course Description: 262A, 262B

Course Objectives

Course Materials

Office Hours

Grading

Summary of Assignment due dates

Topics, Readings, and Guest Speakers

Week 1, April 1: Selection and appraisal

Week 2, April 8: Data management in the field/data management in the university

Week 3, April 15: Technologies of persistence and identification

Week 4, April 22: Data, Publishing, and Publishers: 8am-11:20am session

Week 5, April 29: Economics of data: preservation, access, and sustainability

Week 6, May 6: Provenance in data: WWW and archival approaches, Paul Groth, Free University of Amsterdam and Elsevier Research, by video

Week 7, May 13: Ethics in use of public data for research Prof Katie Shilton, U of Maryland, and Jeffrey Burke, UCLA

Week 8, May 20: Intellectual property in data: Peter Hirtle, Berkman Center, Harvard, by video

Week 9: choice of two topics:

Week 9, May 25: University-Industry partnerships: Data transfer

Week 9, May 25: Data in public engagement

Week 10, June 3: Project presentations

Week 11 (Exam Week), June 8 (Monday): Projects due

References

Course Description: 262A, 262B

Data are both process and products of the research enterprise. Increasingly, data are viewed as scholarly products to be managed, shared, and reused. Funding agencies are requiring data management plans as part of grant proposals, journals are requiring the release of data to reviewers and readers alike, and libraries and archives are adding data to their collections. Managing data is a complex process, involving expertise in technology, knowledge organization, information policy, and in the research domain.
These two courses (winter and spring) survey the landscape of data management, practices, services, and policy, including the uses of data in the sciences, social sciences, and humanities; data management practices (e.g., metadata, provenance, technical standards); national and international data policy (e.g., intellectual property, release policies, open access, economics); management of data by research teams, data centers, libraries, and archives; and data curation, preservation, and stewardship. The courses are intended for graduate students in information studies and in any domain that requires the management of research data. By bringing together students from across campus, these seminar courses will engage students in practical, professional, and theoretical challenges in the use and reuse of research data. Assignments include analyses of data archives, data management plans, curating data for a research team, and domain-specific activities. Students will work in teams with UCLA researchers and will make class presentations. This is a two-term course. Students taking Part I (winter) are not required to take Part II (spring), but Part I is pre-requisite to Part II.

**Course Objectives**

1. Students will learn to distinguish among many forms of data and factors that influence their interpretation in different contexts and over time.
2. Students will learn professional criteria for managing, selecting, and appraising data.
3. Students will learn to use and assess data collections, repositories, and services.
4. Students will gain technical skills in managing data in specific research settings.
5. Students will gain a basic knowledge of practices to curate digital data.
6. Students will learn basic principles of public policies for data.

**Course Materials**

All course materials will be posted on or linked from the CCLE site for this course. Enrolled students have access to the site at [http://www.ccle.ucla.edu](http://www.ccle.ucla.edu).

We will refer also to the two books required for 262A: (Borgman, 2015; Ray, 2014)

**Office Hours**

Wednesdays, 1:30-3:30pm (link posted on CCLE); other times by appointment on Tuesdays and Thursdays, as the instructor is traveling intermittently this term. Office hour meetings can be done by video (Skype, FaceTime), if arranged in advance. Send contact request to CLBORGMAN for Skype or [CLBORGMAN@ME.COM](mailto:CLBORGMAN@ME.COM) for FaceTime.

**Grading**

Assignment 1 (individual work) 30%
Term project (team work) 50%
Class participation and analysis of readings 20%

Details of the assignments are provided on separate documents.
Students are expected to complete all assigned 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 20% of your grade. Written assignments are to be submitted electronically to the CCLE site and on paper at the beginning of the class session, as noted. Assignments will be marked down 2 points for each day late. No assignments will be accepted after midnight Monday, June 8.

**Summary of Assignment due dates**

<table>
<thead>
<tr>
<th>Date</th>
<th>Assignment Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 1</td>
<td>Assignment 1 given</td>
</tr>
<tr>
<td>April 1</td>
<td>Term project assignment given</td>
</tr>
<tr>
<td>April 15:</td>
<td>Term project proposal due</td>
</tr>
<tr>
<td>April 23:</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>May 11:</td>
<td>Team Project Report outline due</td>
</tr>
<tr>
<td>May:</td>
<td>Teams to meet with instructor during office hours</td>
</tr>
<tr>
<td>June 3:</td>
<td>Student presentations</td>
</tr>
<tr>
<td>June 8 (Monday):</td>
<td>Final projects due to CCLE</td>
</tr>
</tbody>
</table>

**Topics, Readings, and Guest Speakers**

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. Prepare some talking points as part of your reading and studying.

**April 10: This is the finalized order of topics with all guest speakers confirmed**

**Week 1, April 1: 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?

(Borgman, 2015), Chapter 10, What to Keep and Why
(Faniel & Jacobsen, 2010) How scientists assess data for reuse
(Gutmann et al., 2009) ICPSR and Data-PASS on data preservation
(Harvey, 2008) Appraisal and selection from DCC
(A. A. Goodman et al., 2014) Advice to scientists for what to keep and how
(National Health and Medical Research Council, 2007) Australian code for responsible data management
(Pfeiffenberger & Carlson, 2011) Selecting data for publishing in a journal
(“Strategic Directions: Appraisal Policy,” 2007) US government policy on appraisal of records
**Week 2, April 8: Data management in the field/data management in the university**

Peter Darch and Elizabeth Stephenson

Data are managed during the research process in much different ways than they are managed later, when contributed to data repositories, libraries, or archives. These

**Data management in the field: Peter Darch**

Researchers are responsible for managing their own data when in the field. Multiple factors can complicate management: data collection plans change dynamically, adapting to unanticipated challenges or new possibilities that emerge; equipment can fail, or prove unsuitable in physical conditions, and limited time means scientists prioritize data collection at the expense of recordkeeping and other bureaucratic tasks.

(Barkhuus & Brown, 2012) Social science fieldwork and collaboration  
(Bird, Willoughby, & Frey, 2013) Lab notebooks in the digital era  
(Borgman, Wallis, & Enyedy, 2007) Field research in ecology, and how they manage their data  
(Lindseth & Baker, 2012) Designing a data logger for oceanography

**Data Management in the University: Elizabeth Stephenson**

Once scholars are ready to release their data, they may offer them to a repository, which may or may not accept those datasets. Data collections and repositories take many forms and reside in many places. Libbie Stephenson will provide an introduction to data repositories, to the technical aspects of data ingest, and professional aspects of repository management. Read the two journal articles on repository management and browse the repository links.

**Repositories – Management issues**

*(Kimpton & Morris, 2013; Peer, Green, & Stephenson, 2014)*

**Repositories** - offering services  
UK Data Service http://ukdataservice.ac.uk/  
Inter-university Consortium for Political and Social Research (ICPSR)  
http://www.icpsr.umich.edu  
DataBib http://databib.org/  
Registry of Research Data Repositories http://www.re3data.org/  
OpenContext http://opencontext.org/ and The Digital Archaeological record (tDAR)  
https://www.tdar.org/

**Repositories** - self-deposit  
Dryad http://datadryad.org/  
figshare http://figshare.com/  
DataShare/U.S. http://datashare.ucsf.edu/xtf/search
DataShare/U.K. [http://datashare.is.ed.ac.uk/](http://datashare.is.ed.ac.uk/)

**Repository Systems**
Islandora [http://islandora.ca/](http://islandora.ca/)

**Week 3, April 15: Technologies of persistence and identification**
The technologies associated with data curation and management vary widely by domain, method, and many other factors. These technical overviews will aid you in developing your term project and doing assignment 1. 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:

(“Data Foundation and Terminology (DFT) WG Recommendations,” 2015) RDA; basic terminology and models for data handling

- 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

(“Data Foundation and Terminology (DFT) WG Recommendations,” 2015) RDA; basic terminology and models for data handling

- Research objects:


Recommended:
(Pepe, Mayernik, Borgman, & Van de Sompel, 2010) Object Reuse and Exchange; life cycles of data objects
**Week 4, April 22: Data, Publishing, and Publishers: 8am-11:20am session**

This class session will start at 8am for a live broadcast of Prof. Borgman’s keynote presentation to the Association of Science, Technology, and Medicine Publishers in Washington, DC, followed by a panel discussion. Dr. Darch will host the session at UCLA and lead the remainder of the discussion locally.

Note: These readings changed from the April 1 version of the syllabus

**Required readings:**

(Bourne et al., 2011) Force11 Manifesto on scholarly communication and publishing

(Borgman, 2015, Chapter 9) Credit, attribution, and discovery

(Starr et al., 2015) Data citation for accessibility

(“Brussels Declaration,” 2007) Publishers’ statement on data and publication

(Parsons & Fox, 2013) Is data publication the right metaphor?


**Sites to browse for background:**

(“RDA/WDS Publishing Data IG,” 2015) see summary at [https://docs.google.com/folderview?usp=sharing&id=0B4qnUFYMgSc-eGFJJaGR4NXJpVVk](https://docs.google.com/folderview?usp=sharing&id=0B4qnUFYMgSc-eGFJJaGR4NXJpVVk)

(DataCite, 2013) Consortium

(Thomson Reuters, 2015) Commercial index to data


(Macmillan, 2015) Digital Science, aggregation of multiple ventures from the publisher of Nature


(Force11, 2015) new models of scholarly communication
Recommended:

(Borgman, 2003) The invisible library
(Edwards et al., 2013) Knowledge infrastructures
(Ginsparg, 2011) history of ArXiv

**Week 5, April 29: 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.

Note: please read ahead for the Provenance topic. Today we will finalize questions to ask Dr. Groth next week, so we can send them to him in advance.

(Beagrie & Houghton, 2014) Value of data sharing and curation
(Berman et al., 2010) US policy study on economics of digital preservation
(David, 2009) Commons and anti-commons for access to research information
(Hess & Ostrom, 2007) Overview of commons models of information
(Kelty, 2012) open access, open science, and motivations of scholars; free riders
(Rosenthal, 2010) Cost analysis for digital preservation of data; keynote presented as long blog post

**Week 6, May 6: Provenance in data: WWW and archival approaches, Paul Groth, Free University of Amsterdam and Elsevier Research, 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.

Note: please read ahead for the Intellectual Property topic. Today we will finalize questions to ask Peter Hirtle, so we can send them to him in advance.

Note: Please read ahead for the options for Week 9 topics so that we can finalize our choice and prepare accordingly.

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

**Week 7, May 13: Ethics in use of public data for research Prof Katie Shilton, U of Maryland, and Jeffrey Burke, UCLA**

Prof. Shilton, a PhD graduate of the UCLA Information Studies program, and Prof. Jeffrey Burke of the UCLA Design and Media Arts, are conducted NSF-funded research on Named Data Networking and policy impacts thereof. They focus today on research ethics and open data.

(Boyd & Crawford, 2012) Critical questions about the use of data in research

(Kramer, Guillory, & Hancock, 2014) Notorious Facebook study and responses in media

(Goel, 2014a, 2014b; Sullivan, 2014)

(Shilton, 2012) Participatory research data in the information sciences

(Zimmer, 2010). Ethics of research on Facebook

**Week 8, May 20: Intellectual property in data: 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.


**Week 9: choice of two topics:**

**Week 9, May 25: University-Industry partnerships: Data transfer**

University-industry partnerships that involve the transfer of technologies developed in university research for commercialization are increasingly common. Central to the success of these transfers is successful exchange of data used in the development and testing of these technologies. This exchange typically involves negotiations, which can include: issues of intellectual property and ownership, protection of research subjects; and data standards and formats.
Reading:
(Evans, 2010)
(Mowery, 2004) – Chapter 5 (pdf available online)
(UCLA Office of Intellectual Property and Industry Sponsored Research, 2014)
(Louis, Jones, Anderson, Blumenthal, & Campbell, 2001)

Two short videos:
(UCLA Technology Transfer and Startups, 2014),
https://www.youtube.com/watch?v=Idun5eYjvS4
(UCLA Inventor, 2014),
https://www.youtube.com/watch?v=nZymuXOrC5Q&feature=youtube_gdata_player

**Week 9, May 25: Data in public engagement**

Libraries, archives, and academic researchers face increasing pressure to engage diverse audiences beyond the university. They may make special collections accessible either physically or virtually, or recruit members of the public to participate in citizen science projects. These processes of public engagement involve selection of research data or artefacts, transforming these research objects accessible into more accessible objects (such as visualizations), and training the public in handling, processing, or interpreting these objects.

Reading:
(Darch, 2014)
(Fortson et al., 2011)
(Raddick et al., 2013)
(Causer & Wallace, 2012)
(Lang & Rio-Ross, 2011)
(Prestopnik & Crowston, 2011)

**Week 10, June 3: 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 8 (Monday): Projects due**
Final projects are due by midnight on Monday, June 8.
References


UCLA Inventor: Dr. Patricia Ganz. (2014). Retrieved from https://www.youtube.com/watch?v=nZymuXOrC5Q&feature=youtube_gdata_player


