

ITPB Subcommittee on Research and Education Data Management

First meeting, Monday, November 26, 2007, 12-1pm, Murphy A244

Attending: Christine Borgman (Chair), Sam Morabito, Margo Reveil, Alan Robinson, Gary Strong

Absent: Marilyn Raphael, Libbie Stephenson

Organized by Joanne Fife

DRAFT minutes by Christine Borgman

We reviewed the slides from the November 1 ITPB meeting which laid out the agenda and charge for the subcommittee (these slides also were sent to the subcommittee). Our agenda was to address the action plan laid out therein:

Campus research and educational data management - i.e., data management as infrastructure:

1. help for researchers to create data management plans
2. develop an institutional response to the directions of grant agencies regarding data
3. policy and standards for setting up databases
4. who owns the problem

We agreed that the campus should take a leadership position on data management for research and education, rather than a reactive position to national and international policy.

We identified a number of areas where we should gather more information about the state-of-the-art around UCLA to understand the scope and scale of the problem. These include:

- Policies on what to keep or curate and not to keep that originate in the library system and are associated with policies for the SRLF and NRLF. It was noted that these facilities are funded as part of UC-wide infrastructure.
- Inventory of data repositories currently on campus.
- Inventory of campus data practices. Some individuals, research groups, and departments have their own repositories; some participate in national or international efforts (Art Toga in Medicine; Paul Davis in Geosciences/seismology); others have no home for their data.
- Studies of data capture and storage practices, such as those done of CENS and CNSI (Borgman et al; Reveil).
- Digital repositories housed by the UCLA library system. Most of these are organized for educational purposes and most are small, but offer important use

cases. They include data for the Center for Digital Humanities and for the Cuneiform Digital Library.

- Inventory of data management plans. OCGA examines data management plans in NIH and other grant proposals. They have developed a template (contact Linda Lee?). Identify DM plans from NIH, NSF, Wellcome Trust, other sources.
- Data warehousing efforts at UCLA and Disaster Recovery Plan (done by IBM). Both have criteria, standards, and plans for data management.

General discussion:

Print preservation is not an adequate analogy for the digital preservation problem. Much of our print is crumbling. As we digitize for preservation purposes, it becomes a digital preservation problem.

Data warehousing is very expensive. Related to this is disaster recovery of which the IBM plan estimates a preliminary price tag of \$14- \$17million over a five year period.

Goals:

The highest priority for action is item #1, which should be extended to include education, i.e., *help for researchers and educators to create data management plans*. It is likely that considerable reinvention is taking place that could be avoided through better campus coordination.

Responses to item #2 may emerge from study of #1.

Policies and standards for setting up databases are insufficient. We also need the means to store massive amounts of data. The infrastructure for data storage is very costly.

Who owns the problem is the biggest unanswered question, and is unanswerable at this time. We must chip away at it in parts.

Next steps:

Obtain staff resources to start gathering some of these records and policy documents from around campus.

Distribute to committee.

Convene full committee by middle of winter term to review these materials.

ITPB Subcommittee on Research and Education Data Management

Comments from Libbie Stephenson, Director, ISSR Data Archive

I am not sure what all has been covered thus far so forgive any questions on things already hashed out.

Data – has there been a definition of what we mean when we say “data”? In different disciplines it can have different meanings. In the social sciences we think of “data” as data sets containing numerical representations to survey responses. These are electronic files, often in ascii format. Other formats in other research disciplines can include images, audio, text, moving images, databases with front end software, and so on. Each format has different needs in terms of long term management.

Campus data practices – In the social sciences, much has been done in a multi-institutional collaborative approach with an emphasis on open source/open access solutions. UCLA is a member of the Inter-University Consortium for Political and Social Research (ICPSR) <http://www.icpsr.umich.edu>. Member institutions have agreed to manage data locally and at ICPSR according to agreed upon standards.

Data management plans – for the social sciences see: <http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf>; ISSR provides support to grant applicants and others in the College (actually, we do it for anyone on campus involved in social/behavioral research) to develop data management plans in proposals, in accordance with the requirements of the funding agencies, and for depositing datasets either at UCLA or other depositories. This includes public and restricted access data sets.

Curation and preservation – in the social sciences a metadata standard used to describe and preserve data files throughout the data life cycle has been developed and is in use here at UCLA. ISSR assists researchers (faculty and grad students) by using the standard as one part of our preservation program. See the Data Documentation Initiative: <http://www.ddalliance.org/> The ISSR Data Archive uses migration and refreshing to ensure the longevity and usability of the files.

Warehousing – this is where I see one of the challenges – in the social sciences the amount of storage space needed is small compared to the sciences. Image, audio and moving image files are also large. I would encourage a multi-institutional collaborative open source/open access set of solutions.

Educating the researcher – faculty and graduate students need to know how to manage their data and how to use tools available to them. They need to know who is available to assist them. Research data management needs to take place

throughout the entire life cycle of the research but most wait until the end of a project; this can be costly. They need educating (still) on the policies and regulations that affect data-related research.

To add to the inventory: Communication Studies houses a News Archive of audio and moving image files of news programs. They seek a home and support. Contact Paul Rosenthal (prosenthal@commstds.ucla.edu)

Who owns the problem? UC Policy addresses part of this as follows:

General University Policy Regarding Academic Appointees

II.5. Publicity of Results

"Notebooks and other original records of research are the property of the University ."

This policy was written in 1958 and has not been modified. See:
<http://www.ucop.edu/acadadv/acadpers/apm/apm-020.pdf> (page 4)

In the social sciences the culture and practice has been to permit faculty to retain control over their data, even when the data was collected using outside funding. Faculty have been permitted to take data with them when they leave the UC, to deposit or not, to share or not and to set their own time frame after which they will share/deposit their data. They also have the choice of where to deposit data; they do not have to ensure a copy is physically stored at UCLA. They can also manage their own data distribution through a web site or similar open access. It might be worth understanding the data culture in other disciplines. In different research areas, long term storage or preservation of data has not been widely adopted. Solutions to the "problem" should probably reflect these cultures.