Report of the Restricted Use Data Task Force

Background and Summary of Charge

In August 2013 the Library Administration via AUL Beth Dupuis, and UC Berkeley D-Lab Acting Director Cathryn Carson charged a group of librarians and D-Lab staff to investigate issues surrounding the acquisition, licensing, access, and curation of restricted use data (defined below) for the UC Berkeley campus. As noted in the charge, there is a sustained and growing interest in obtaining and providing access to restricted use data among campus departments, including but not limited to Economics, Sociology, Agricultural Economics, Demography, Public Health, Public Policy, and Business. The task force was charged with exploring models from peer institutions, consulting with campus researchers and colleagues to ensure that recommendations align well with research needs and workflows, and recommending next steps.

Restricted Use Data Defined

Restricted use data (RUD) may be defined in this context as data where the provider does not permit unrestricted access by the public or the university community. RUD differs from "public use" or aggregate data where identifiable information from survey respondents has been coded or deleted to protect confidentiality. Typically, RUD must be securely hosted and protected from unauthorized access or disclosure in accordance with the terms of the license. Usage restrictions, access, and review of resulting research may be specified in a set of security procedures. The degree of restriction may differ substantially, depending on the terms of the license, which range from simple one paragraph agreements to extensive contracts requiring signatures from high-level campus authorities.

Current Campus Practices and Patterns

At present most campus researchers purchase and secure RUD independently or collaboratively within departmental silos. This siloing can create problematic inconsistencies and inefficiencies. Some departments simply sign off on agreements and leave it to the individual to manage. Researchers across campus may also pay for the same data several times since others may not know the data has been secured. Researchers who already know what to do and who to consult with on campus negotiate access more easily, but few people have accurate and complete knowledge of the process. This places many researchers at a disadvantage and the campus at risk. By taking leadership in this area, the Library and D-Lab will provide greater and more equitable access to these scarce information sources, enhance the ability of researchers to produce significant work, and provide greater legal protection for the campus by ensuring the data is licensed, used, and curated in a systematic and responsible fashion.

Campus Stakeholders

The task force met a total of five times and invited Julie Conner, Associate Campus Counsel at the UC Berkeley Office of Legal Affairs, to advise us. We also met with Associate University Librarian Erik Mitchell in his role as Director of Digital Initiatives & Collaborative Services, Lynne Grigsby (Library Systems Office), Lydia Peterson (Library Licensing), and Lisa Rowlinson de Ortiz (Head of Cataloging) to discuss acquisitions, access, and licensing work flows. It should also be noted that almost 100 economics graduate students and faculty signed a petition urging action on this in Spring 2013. There are also potential synergies with other campus partners such as the iSchool, BRCOE, or the Special Projects Office. Two campus partners, the Office for Protection of Human
Subjects (OPHS) and Information Services and Technology (IST) are particularly important. OPHS administratively supports the Institutional Review Boards (IRBs) which ensure that human subjects of research are protected, while IST supports campus computing infrastructure, and maintains secure facilities for server-based research activities. Both partners have been working with D-Lab to build out training and access for researchers, as well as the development of models and best practices suitable for RUD.

**The Need for Partnership**

D-Lab and the Library have been working out their respective roles in providing data services to campus users. The model we wish to embrace will be collaborative, building on each unit’s strengths and expertise. The Library excels in providing access and effective creating discovery tools for information resources (descriptive metadata, research guides, and API generated lists of information sources). The Library has a collections budget and departmental liaisons who work with faculty and grad students who have an interest in acquiring numeric data. The Library also has a 50% time license librarian with previous experience licensing RUD contracts at a peer institution (Harvard Business School Library).

D-Lab staff have extensive experience hosting RUD in restricted facilities like the California Census Research Data Center, which provides access to non-public data from the Census Bureau and other federal agencies, and provisioning ‘cold rooms’ and secure environments for researcher access to other datasets that do not fit the high-security CCRDC framework. The D-Lab is familiar with understanding and complying with security requirements imposed by providers, and in assisting researchers with processes and paperwork needed to obtain access.

**Models for Acquisition and Access**

In the areas of acquisition and access the potential for library and D-Lab collaboration is particularly strong. RUD may be either free or priced, and may be a one time or ongoing acquisition. In some instances the data may need to be destroyed upon the completion of a research project. The following scenarios identify possible models for addressing each of these variations.

a. **Priced RUD, one-time purchase**

Subject librarian coordinates with Jon Stiles about the initial license review. Decide on usage restrictions and create a data management and security plan to be attached to the agreement, and assist researchers with required paperwork. If a decision is made to purchase, use the Library Materials Acquired Digitally (MAD) process ([http://www.lib.berkeley.edu/Staff/CS/dt_home.html](http://www.lib.berkeley.edu/Staff/CS/dt_home.html)). Consult with the licensing librarian (Lydia Peterson) and data cataloger (Trina Pundurs) via the Footprints process. Data file will be cataloged in Oskicat, listed on library research guides, and potentially scoped in the Electronic Resources Finder (ERF). As appropriate, assist researchers with creating and using a local secure environment which complies with use agreements.

**Example:** Annual Survey of Industries, NSSO India (See Case Study 1)

b. **Priced RUD, subscription**

Subject librarian coordinates with Jon Stiles at the D-Lab for the initial license review. Decide on appropriate restrictions on use, and create a data management plan to be attached to the agreement. If a decision is made to purchase, use the Library Electronic Resources Order form ([http://ucblibrary3.berkeley.edu:8080/emfo/triage](http://ucblibrary3.berkeley.edu:8080/emfo/triage)). Advise and consult with the licensing librarian
(Lydia Peterson) about licensing via the Footprints process. Data file will be cataloged in Oskicat, recorded on library research guides and potentially scoped in the ERF. In most cases, subscriptions for RUD would also be housed in a secure location and would not be accessible via IP recognition. Password access is a possibility.

**Example:** *Nielson Data Sets* (See Case Study 2)

c. ‘Free’ Restricted Use Data

There are examples of RUD that can be obtained without money costs. The licensing and usage restrictions for these vary. Some require signatures from local campus research units (e.g. a department or center) while others require higher level campus approval. After some discussion, the committee recommends licensing for these products be negotiated and discussed between the library and D-Lab. If the library were asked first, the steps would be similar to the above: contact Jon Stiles for the initial review of the license, decide on appropriate restrictions on use, and create an internal data management plan to be attached to the license. Access could be via Oskicat, Library Research Guides or through the Electronic Resource Finder (ERF). The Materials Acquired Digitally (MAD) order form could be used, but the acquisitions process could also be undertaken independently by the D-Lab in the absence of money costs.

**Example:** *UN Gender and Generations Programme Survey* (See Case Study 3).

d. Limited Time Use Data

There are examples of RUD the campus may wish to acquire that must be destroyed after conclusion of the project. The D-Lab has experience working with these data, but the library is not well positioned to purchase or catalog data that will be destroyed after use. However, the library may be willing to assist the D-Lab by facilitating licensing, acquisition, and access using the steps outlined above. For example, the library could help negotiate license terms and create entries on research guides indicating the steps needed to access the data. It is likely that the same data sets will be acquired by multiple users over time.

**Example:** *National Longitudinal Study of Youth* [http://www.bls.gov/nls/nlsdata.htm](http://www.bls.gov/nls/nlsdata.htm)

e. Locally Held Data

There are examples of RUD that the campus may have developed locally or which the University has long-standing use-rights to, but which require protection due to confidentiality concerns. The library would not be positioned to acquire that data, but could participate in cataloguing it and expanding knowledge and access to the collection.


As an additional note, the D-Lab also has California Field Polls and historic international datasets for which there is currently limited or no public metadata. Cataloging these datasets in Oskicat, and additional uploading of datasets into Oskicat in the future, would enhance access to this information. The committee notes that the metadata standard for numeric data (DDI) has been cross-walked into MARC both at UCB and nationally.

**Budget and Inventories**

Per our conversations with Lydia Peterson, Harvard University reserved $100,000 annually for
RUD. UC Berkeley departments here have not communicated this degree of interest, but the need for additional funding for numeric data is large. (The D-Lab already provides $140,000 in funding for the CCRDC.) In the area of numeric data in general, UC Berkeley does not have access to many ‘unrestricted’ databases and data sets our peer institutions have acquired. An all-too-common phenomenon here is faculty who arrive expecting support from the library in this area, do not find it, and procure the data individually or travel elsewhere to obtain it.

**Licensing Models**

The D-Lab has extensive experience negotiating RUD agreements and management plans. The library is fortunate to have a new librarian with RUD licensing experience, and envisions a future licensing process that will be mutual, particularly for priced data sets. The committee wishes to stress the importance of such a skill set to the library, and recommends this position be permanently retained to perform this function.

On January 21, 2014 the task force met with Julie Connor from the Berkeley Office of Legal Affairs. It was Julie’s opinion that liability would reside either with individuals who sign data agreements, or with UC Berkeley as an institution, not with either the Library or D-Lab. Risk would be mitigated if D-Lab takes on the responsibility of monitoring compliance with the agreements. Julie called D-Lab’s monitoring role “crucial.” She also suggested that the question of ownership should be stated explicitly in the agreements with the data providers. Any restrictions on the use or the users of the data would have to be noted in the publicly-visible catalog record. Julie also suggested the Library assign a single person to handle the acquisition of restricted use data, and recommended the development of a form or checklist to be consulted.

The task force suggests applicable licenses be signed jointly by the library and the D-Lab. If the license requires special security provisions such as individual supervision, research review, or submission of resulting publications to the data provider, the task force recommends that D-Lab take responsibility for these functions, and that these be recorded in a data management plan. The task force also discussed potential blanket restrictions for types of data the campus may not wish to acquire. This might include data not available to all campus users (e.g. restricted to graduate students and faculty) although there was not complete consensus on this.

**Models for Curation, Security, and Research Review**

By definition, access to RUD is “gated” and monitored. It is expected the primary audience for these data will be faculty and graduate students. In some cases, D-Lab and the library will act directly as gate-keepers, and in other cases they will help steer and support researchers in agreements with data providers and local institutions like the Office for the Protection of Human Subjects (OPHS).

It is expected that one of the primary roles undertaken by D-Lab in this initiative will be to house RUD in secure location(s). In some cases access could be granted after an appropriate interview has been conducted; in other cases official registration with the data provider will be required in addition to clearance from the Institutional Review Board (IRB). The degree of the restriction will be the joint decision of D-Lab, researchers, and data providers, but the classic example is a secure “cold room” with no access to the Internet or external ports to download data. In cases where researchers need to register with the data provider, compliance will be supported and documented by the D-Lab.

One of the outstanding questions the task force faced was how shared responsibility and service points for these data collections would be managed. The Library will own some data for which D-
Lab will be the service point. The task force explored the following options for partnership:

- Make the D-Lab an official library location, similar to an affiliated or departmental library. This would create an official location in Oskicat indicating where the data would be housed. This would also require creating a location code, adding D-Lab information to the library hours database, and having the D-Lab supply the library with current contact information.

- Make D-Lab and the Library Data Lab a "shared subject specialty library" with a web page that addresses the collaboration and the project. The page could incorporate an ERF list of data resources (see below).

- Use the ERF software to list RUD sources with notes about restrictions and the location. This would involve scoping the ERF so RUD would be its own media type.

Of these three options, the task force supports the first in combination with the third. It is very desirable to create permanent quality metadata using Marc records. The second option would be easier to implement but would create an extra step of referral. The third option we see as an excellent supplement to but not a replacement for the first option, which is also strongly supported by the interim director of the D-Lab.

**Policies at Peer Institutions**

Acquiring and providing access to RUD offers the library and campus an opportunity to provide leadership and innovation in an emerging field. The following are examples of the growing number of peer institutions where data centers and/or libraries provide services for acquiring and accessing RUD. The total number is much more widespread,

- **Princeton University Library.** Licenses for data are signed and negotiated by the economics librarian and statistical consultants. The contracts include those for data requiring registration and signatures by users. Data is housed on secure servers. Access is overseen and research is reviewed by statistical consultants. Selected examples include:
  
  - Israel Census 1995
  - National Longitudinal Survey of labor market experience
  - Northern Ireland Sample of Anonymised Records
  - Public Patient Discharge Study (California)
  - 1991 Pakistan Integrated Household Survey

- **Harvard Business School Library.** Annual budget of $100,000 for RUD. At least five faculty must show interest. The library keeps records all transactions and negotiates the license. The library does not purchase data required to be destroyed on completion of the project although it does help with licensing. In most cases these data are purchased for groups of faculty and housed on faculty machines.

- **Johns Hopkins University Library**
  [http://guides.library.jhu.edu/content.php?pid=62430&sid=1883583](http://guides.library.jhu.edu/content.php?pid=62430&sid=1883583)

- **University of Michigan Population Studies Center**
  [http://www.psc.isr.umich.edu/dis/data/restricted/r-data.html](http://www.psc.isr.umich.edu/dis/data/restricted/r-data.html)
Memorandum of Understanding (MOU)

The task force recommends an MOU with respect to ownership, access, licensing and usage be drafted. The MOU should stipulate any data purchased by the library and housed in the D-Lab will be jointly owned. The MOU should state that adequate precautions will be taken to ensure long term preservation, access, and compliance with signed license agreements. It should also provide a mechanism for the library to reclaim ownership of the data in the event of D-Lab closure, relocation, or unwillingness in future to abide by the terms of the agreement. Julie Connor has agreed to review the MOU after it has been drafted.

Conclusion

In the past several years the library and D-Lab have rejected numerous requests from faculty and graduate students to acquire and provide assistance with RUD. The case studies outlined below (see appendix) provide specific examples. It is very clear to the task force that this is detrimental to the interests of all parties, and that the use of RUD is an emerging research trend that must be addressed soon. The RUD Task Force recommends as next steps:

- Drafting an MOU
- Procuring a budget for data
- Initiating a pilot project to test the model.
- Providing a publicly viewable series of steps on the web, such as those published by peer institutions.

We are grateful to library administration and Cathryn Carson in charging this group to address this important need and look forward to its successful resolution.

Respectfully Submitted,

Jon Stiles, Interim Executive Director, D-Lab
Jim Church, Librarian for Economics and International Government Information
Elliott Smith, Bioscience & Emerging Technologies Librarian
Mary Elings, Archivist for Digital Collections, The Bancroft Library
Introduction

The Annual Survey of Industries (ASI) is the principal source of industrial statistics in India. It provides statistical information to assess and evaluate changes in the growth, composition and structure of organized Indian manufacturing sectors. The Industrial sector occupies an important position in the Indian economy and has a pivotal role to play in India's economic development. The collection and dissemination of ASI data on a regular basis are thus of vital importance. The Survey is conducted annually. Each round of the data costs a few thousand dollars.

The Need at UC Berkeley

This dataset and others distributed by the National Sample Survey Office (NSSO) of India have been requested by faculty and graduate students in economics more than any other over the years. The obstacle to access has been the following undertaking, quoted in part below and available at http://mospi.nic.in/mospi_new/upload/undertaking.pdf.

(i) The confidentiality of the unit level data will be maintained and adequate precautions would be taken for not disclosing the identity of the units directly or indirectly.

(ii) The data would be used after understanding the concepts, definitions, design and coverage of the survey for a proper appreciation of the limitations and nature of the data and for obtaining meaningful estimates and results.

(iii) The data obtained as above will not be passed on either wholly or partially with or without profit to any other data user or disseminator of data with or without commercial purpose.

(iv) The data user shall acknowledge the data source in the research output.

Signature..............................................

Date.................. Name..............................................
Case Study 2: Nielsen Data, University of Chicago Business School
http://research.chicagobooth.edu/kilts/marketing-databases/nielsen

Introduction

These datasets, including both Consumer Panel Data and Retail Scanner data, were recently made available to researchers nationwide. The Consumer Panel Data provide a representative panel of households that continually provide information about their purchases in a longitudinal study. Participants use in-home scanners to record all of their purchases intended for personal, in-home use. The Retail Scanner data consist of weekly pricing, volume, and store environment information generated by point-of-sale systems from more than 90 participating retail chains across all US markets. This is the kind of dataset that lends itself to the sorts of "big data" analysis currently attracting a great deal of attention in libraries and beyond.

The Need at UC Berkeley

In August 2013, an assistant professor of Economics (Benjamin Faber) wrote to Jim Church hoping the library could acquire a campus-wide license for this product. The cost was approximately $5,500 and several faculty offered to help the library co-invest in it. The license required that data be kept secure while in use, that all campus researchers register with the provider, and that users supply to Nielsen any resulting publications, among other restrictions.

Upon showing the agreement to the licensing librarian at that time amidst discussion for this as a potential pilot, the proposal was rejected, leaving several faculty feeling frustrated. With the library out of the picture the economics department worked to procure a campus-wide license to the product on their own, using faculty research funds. The license and access to the data are now administered by a single staff member in the department.

Upon reflection it seems clear the library should have been better equipped to address this need and handled the situation differently. Although through dogged efforts by faculty members, access to these data was eventually negotiated, the library missed an excellent opportunity to demonstrate responsiveness and innovation to the campus. While libraries have traditionally not handled this kind of product, these data acquisition efforts are a part of the support our faculty and students are now looking to the Library for, and which the Library and D-Lab have the capacity to deliver.
Case Study 3: UN Gender and Generations Programme Survey
http://www.ggp-i.org/data/data-access.html

Introduction

This a longitudinal survey of 18-79 year-olds in 19 countries that aims to improve understanding of the factors which affect the relationships between parents and children (generations) and between partners (gender). A broad array of topics including fertility, partnership, and the transition to adulthood, economic activity, care duties and attitudes are covered. At the time of library involvement the data was freely available but required the signature of a department head and a secure storage facility to house the data.

The Need at UC Berkeley

The student requesting the data was in the economics department, although these data have utility for other campus researchers. In this case the prospective researcher was an undergraduate who searched broadly across campus (including asking the chair of the economics department) for a signature and a location for the data on non-networked, password protected computers. At length the request was referred to Jim Church, who referred it to Jon Stiles, who referred it to the campus Demography Lab which has its own cold room. A staff member at the Demography Lab signed the license for the grateful undergraduate. There the data still sits. There is no official library access point and the existence of this dataset (like many others on campus) exists only in the heads and emails of those in the know.
Workflows and Responsibilities

Researcher communicates needs to D-Lab

D-Lab

D-Lab and The Library communicate data access, resources, and supports to researcher

Researcher communicates needs to Library

The Library

D-Lab and The Library coordinate and communicate routinely

Dataset Request: Restricted?

No

Purchase or license as appropriate, catalogue, preserve and curate
Purchase or license as appropriate, catalogue, preserve and curate locally
Purchase or license as appropriate, catalogue, preserve and curate locally

Determine Access Requirements
- How restricted
- Necessary signatures
- Costs and responsibilities
- Environment

Yes

Signatures, Contracts/Licensing
Assistance and Approvals, IRB
License Agreement w/provider
Creation and placement in “right” environment
Catalog
Manage and Monitor

No