Patron Driven Data Acquisitions: Prizes, Perils, and Pitfalls

Jim Church & Josh Quan
Beyond the Numbers: The Economic Data Ecosystem
Federal Reserve Bank of St. Louis, November 2018

Background

- **Faculty and students requesting data sets from librarians for years, particularly in economics.**

- **Requests for specialized data sets (robots, gasoline, cryptocurrencies, rental prices).**

- **Requests for restricted or "sensitive" data requiring secure storage and research review.**

- **Data is expensive and cost prohibitive for most graduates and undergraduates.**

- **No data, no research. You can't interlibrary loan a data set.**

The Assumptions

Faculty and graduate students typically know what data they need.

As professional researchers, often better than librarians do.

They generally do not approach librarians until other options are exhausted.

The Pre-DAAP Landscape

- **Students and faculty sent librarians requests for data.**

- **Librarians scrambled and horse-traded to get the funds.**

- **Swaps and deals. Joint funding with departments.**

- **Special one-time funding debates at selector meetings.**

Questions

- **Why should libraries collect data sets?**

- **Don't we have it already? Did you look in...?**

- **Shouldn't all data be open access?**

- **Why should we spend money on just a few researchers?**

Aspirations

"James Church (2008) describes the challenges of acquiring international survey microdata, much of which must be purchased and is in high demand by researchers, and recommends strategic data acquisition in consultation with researchers. He advocates for a coordinated, user-driven acquisitions model that includes data within its purview, stating "*we buy many books, documents, and microfiche that few people use…it seems misguided to spend thousands of dollars buying items students have not specifically requested while neglecting to purchase data that could lead to the publication of a doctoral thesis*" (p. 16). Church describes one particular type of data and user need, but these principles can be extended to many disciplines and data types."

-- "Assessing a Patron-Driven, Library-Funded Data Purchase Program." T*he Journal of academic librarianship* Sheehan, Beth yr:2017 vol:43 iss:1 pg:49.

The Road to DAPP

- **Librarians and campus D-Lab partners make the case for acquiring restricted use data and patron initiated data requests.**

- **Economics grad students distribute letter and a petition.**

- **Director of the D-Lab liaises with the Library.**

- **The library forms the first of many data working groups.**

- **A central data fund set up for one time purchases.**

The Illinois Data Purchase Program

Library/D-Lab Working Groups

Berkeley
UNIVERSITY OF CALIFORNIA

Data Sharing, Storage and the Social Sciences Working Group

Recommendations and Report

December 9th, 2016

Working Group Membership:
Jamie Wittenberg (Chair)
Jim Church
Harrison Dekker
Celia Emmelhainz
Jon Stiles

Data Purchase Working Group
PRELIMINARY REPORT

FEBRUARY 2015

HARRISON DEKKER, CHAIR
JIM CHURCH
LYNNE GRIGSBY
LYDIA PETERSEN
SUSAN POWELL
TRINA PUNDURS
HILARY SCHIRALDI
ELLIOTT SMITH

D-Lab/Library Restricted Use Data Partnership

Planning Considerations

- **Eligibility**
- **Licensing/Acquisition**
- **Systems Support**
- **Discovery & Cataloging**
- **Storage**
- **Publicity**

Initial Eligibility Considerations

Numeric, geospatial, and textual data will all be eligible. Funding for data memberships and for software in support of data services will also be supported.

Restricted use data sets and data sets available to only select campus users are eligible for funding. Priority will be given to data sets available to all UCB students, faculty, and staff.

Cost sharing is encouraged, and may increase the likelihood of a proposal being funded.

Grants awarded quarterly and limited to $5000. Emphasis on multi-disciplinary data sets.

Publications must contain the following acknowledgment: "The author wishes to acknowledge the Library at the University of California, Berkeley, which provided support for the completion of this research."

Eligibility in Practice

Due to demand, grants were eventually awarded on an ongoing basis.

Regular meetings to consider grants were abandoned.

The initial cost limit ($5000) was an underestimate. The average cost was over $8000 and several data sets were considerably more.

Funding was transferred from central collections to the data services librarian.

Acquisitions

- **Typical Scenario - Grad Student or faculty email a librarian to inquire about a data source.**

- **Librarian meets with the researcher to ascertain eligibility, costs, and relevance.**

- **Potential cost sharing options discussed with colleagues and departments.**

- **Librarian and/or graduate student contacts the vendor to make inquiries.**

- **Librarian Refers researcher to the [Data Acquisition and Access Program web page](#).**

- **Researcher submits the proposal using the [Library Purchase Recommendation Form](#).**

Data Acquisitions Program Web Site

Purchase Recommendation Form

Materials Acquired Digitally (MAD)

Electronic Resource Order Request Form
(purchased and remotely hosted)

Please provide as much of the following information as possible. An asterisk (*) denotes a required field

Name: James A. CHURCH
Email: jchurch@library.berkeley.edu
Campus Phone: +1 510 768-7611

*1. Title of resource:
*2. Publisher/Provider:

*3. Publisher Contact Name:

*4. Type of resource (Select one; enter details in the "Other" box below):

E-journal (single e-journal title)
- ○ Online only
- ○ Online only, cancel existing print order
- ○ Add online to existing print order
- ○ New print + online

E-journal package (several e-journal titles)
- ○ E-journal package
- ○ E-journal package, with cancellation of print titles. *If applicable, include details in the "Other" box below*

Journal backfile/archive
- ○ Single title backfile
- ○ Backfile of multiple titles

Ebook(s)
- ○ Single title
- ○ Collection or package (multiple titles)

Other
- ○ Please describe below

Other resource type information

Vendor Negotiations

- **This can be time consuming - weeks and even longer are not unusual.**

- **Some commercial data vendors have little to no experience with libraries.**

- **Licensing may be inappropriately restricted (or relaxed). Initial review of licenses by the librarian is important.**

- **Negotiations may involve differences in communication styles and legal environments.**

Licensing

Data licensing can also be both complex and time consuming.

Many data providers are unaccustomed to dealing with libraries:

- Data restricted by department
- Data only for faculty and Ph.Ds
- Data requiring a signature and account for each user
- Data that needs to be returned/destroyed after license expires

Our licensing librarian worked for extended periods on some licenses.

Ticketing

"Every stage of the electronic resource life cycle, from trial, to acquisition, to activation, to maintenance and troubleshooting, to renewal or cancelation, requires a complex series of nonlinear tasks, typically involving hand-offs between multiple library staff members and follow-ups with external players such as students, faculty members, administrative offices on campus, and/or vendors."

"Librarians and staff nevertheless must find ways to complete the work. Often, we struggle through our efforts using a combination of emails and shared spreadsheets. This can often lead to communication issues due to a lack of transparency, poor service outcomes and frustration as issues reported via emails get buried in staff inboxes, confusion as to whom has completed what tasks, and lack of documentation."

-- Smith, Kelly. "Problem Solved! Managing Electronic Resource Workflows Using Ticketing System Software," *Serials Review*, V42:1, (2016). http://dx.doi.org/10.1080/00987913.2015.1137674

Graduate Students

Careers and/or degrees can be determined by the success or failure of a data project.

Grad students & fadculty have domain expertise to identify needs and issues.

Use caution in allowing grad students to interact with vendors.

Ensure students understand usage provisions, e.g. not taking data with them after they graduate or sharing restricted data.

Undergraduates

Some of the most successful research at UC Berkeley conducted by undergraduates.

Limiting proposals to faculty & grad students raises equity questions.

Undergraduates have shorter time constraints. Make sure they understand failure is a possibility and the process can take time.

Selector Due Diligence

Don't forget documentation

Don't forget about file sizes & formats

Don't forget to get data samples

Consider all aspects of confidentiality

Anticipate technical/computing constraints

Systems Support

- **File sizes, data structures, and transfer speeds are considerations that should be addressed by selectors & liaisons.**

- **IT not equipped for making large (>50Gb) datasets available to researchers easily.**

- **Complicated further by large hierarchical datasets of varying quality (XML).**

- **The responsibility for understanding the data and what is expected upon delivery falls to the selector and requesting academic.**

Ticketing and Communications

- Current workflow uses Footprints, a proprietary + legacy system designed for tracking events during the acquisitions process.

- CORAL, an open-source electronic resource management system, is being investigated as a supplemental tool to manage complex licenses, communication, and workflows.

Expectations and Time Constraints

Some data can take weeks or months to process. The longest took several months.

Lost checks, administrative delays, challenging licenses, communications issues.

Multiple parties involved in a nonlinear process: licensing, acquisitions, IT and catalogers.

Students and faculty need to have realistic time expectations.

Computing and Storage Environments
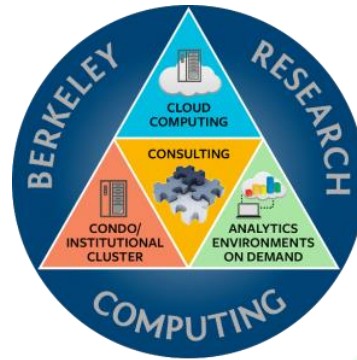
Plans for a Berkeley Dataverse

Cataloging and Metadata

- **Data cataloging is challenging.**

- **Complex workflow issues.**

- **Lack of catalog records for unique data sets.**

- **Neither the cataloger nor selector permitted to view restricted or sensitive data.**

Sample Metadata

Restrictions on Use

- **As required by U.S or National Law (US Census Bureau, EU, Statistics Canada, UK Data Center).**

- **As stipulated by the vendor.**

- **As required by the University (Institutional Research Board, Human Subjects).**

  When in doubt err on the side of caution and consult a licensing or legal expert

Qualtrics & Terms of Use

Berkeley
UNIVERSITY OF CALIFORNIA

Library UNIVERSITY OF CALIFORNIA | D-Lab dlab.berkeley.edu University of California, Berkeley

NYSE EOD summary data, 2007-2016

Need more information about these data?
See: http://dlab.berkeley.edu/data-resources/data#NYSE_EOD

TERMS OF USE
You are applying for access as an authorized user to extract and/or compile data from our locally loaded copies of the archived licensed content, provided that all such activities are within the bounds for fair use and are for the teaching, learning and research purposes. All supplemental terms, including any applicable display and use restrictions continue to apply.

Please read the terms of use below. If you agree to them, click on the "I Agree" button to proceed. If you do not agree, you can click on the "I Do Not Agree" button to return to the home page.

I agree that I will use these datasets solely for non-commercial research, statistical, or educational purposes and NOT for investigation of specific research subjects or commercial gain (i.e. by reselling, redistributing or republishing licensed content).

I agree that I will NOT use the materials in any way prohibited by applicable laws

| I agree | I do NOT Agree |

I agree that my access to these data is contingent on my active affiliation as faculty, staff or student with UC Berkeley, and I will cease to use these data when my affiliation ends.

| I agree | I do NOT agree |

I understand that some uses of the data may require approval of the Berkeley Office of Human Subjects. I have looked at the guidelines. and (select one option below)

My use of the data does not require IRB approval

I have obtained IRB approval for my use of the data

I need to obtain IRB approval, but have not done so yet

| >> |

Publicity

★ **Reach out to faculty and students**

★ **Strike the right balance**

★ **Make time and cost constraints clear**

★ **Do not over-promise**

Lessons Learned

- Some attempts to acquire data ended in failure, at times after long time periods (negotiations)
- Some data did not fully meet the needs of the researchers after the data arrived (due diligence)
- Some data so large (12 Tb) it tooks extended periods to download. Use was problematic due to storage and computing constraints (storage/computing)
- Some licenses took exceptionally long times to finalize (licensing)
- Some data had confidentiality issues for UC Berkeley even though these were not specified in the license (institutional restrictions)
- Some data was "lost" at various points in the process (workflows/ticketing)

Success Stories

Significant Research Data Sets were acquired from the program which offered new research opportunities to grad student and faculty, including:

- **India National Sample Survey Data, Consumer Expenditure Series (Seven Rounds) and Annual Survey of Industries (Four Rounds)**
- **Linguistic Data Consortium (292 data sets)**
- **Nielsen Consumer Panel and Retail Scanner Data**
- **AirBnb Data for Six Metropolitan Areas (2014-2016)**
- **Imports and Exports of Related Parties, US Census Bureau**
- **NYSE Retrac/Protrac data (Firm-level trading of shares by retail investors on the NYSE)**
- **Catalyst (voter registration data)**

AirBnb (AirDNA)

India National Sample Surveys

Title: India - National Sample Survey - Consumer Expenditure Survey Series

Data Use Agreement: < Apply for Access Here >

Description: Surveys of expenditures incurred by households on the consumption of goods and service during the reference period, as well as enterprises owned by households and used by their members during the reference period. Similar surveys are used in the United States and other countries and are used to track consumption and spending patterns of the population and growth patterns for consumer goods.

Detailed metadata can be found at: The international Household Survey Network for:

(Schedule 1.0, Type 1) **1983, 1987-1988, 1993-1994, 1999-2000, 2004-2005, 2005-2006, 2006-2007, 2007-2008, 2009-2010, 2011-2012**

(Schedule 1.0, Type 2) **2009-2010, 2011-2012**

(Schedule 10.0) **1983, 1987-1988, 1993-1994, 1999-2000, 2004-2005, 2009-1010, 2011-2012**

Producers: Central Statistics Office (Industrial Statistics Wing) - Ministry of Statistics and PI, Government of India

Universe: Households and personal enterprises in India, with exceptions for inaccessible areas.

Geographic coverage: India. The survey covers the whole of the Indian Union except (i) interior villages of Nagaland situated beyond five kilometres of the bus route and (ii) villages in Andaman and Nicobar Islands which remain inaccessible throughout the year.

Time Period: 1983 - 2012, Assorted Years, Annual Surveys

Unit of Observation: Households

Smallest Geographic Unit:

Title: India Annual Survey of Industries, 2008-2013

Data Use Agreement: <Apply for Access Here >

Description: The Annual Survey of Industries (ASI) is the principal source of industrial statistics in India. It provides statistical information to assess changes in the growth, composition and structure of organised manufacturing sector comprising activities related to manufacturing processes, repair services, gas and water supply and cold storage. The survey is conducted annually.

Detailed metadata and documentation can be found at the International Household Survey Network for:

2009-2010, 2010-2011, 2011-2012, 2012-2013

Producers: Central Statistics Office (Industrial Statistics Wing) - Ministry of Statistics and PI, Government of India Citation: Smallest Geographic Unit:

Geographic Coverage: India. The geographical coverage of the Annual Survey of Industries, 2008-2009 has been extended to the entire country except the states of Arunachal Pradesh, Mizoram and Sikkim and Union Territory of Lakshadweep.

Time Period: 2009 -2013, Annual

Unit of Observation: The unit of enumeration is the establishment - a factory, workshop, or undertaking - although a consolidated return is permitted for establishments in the same state and industry that share common ownership.

Universe: The survey cover factories registered under Sections 2m(i) and 2m(ii) of the Factories Act, 1948 i.e. those factories employing 10 or more workers using power; and those employing 20 or more workers without using power. The survey also covers bidi and cigar manufacturing establishments registered under the Bidi & Cigar Workers (Conditions of Employment) Act, 1966 with coverage as above. All electricity undertakings engaged in generation, transmission and distribution of electricity registered with the Central Electricity Authority (CEA) were covered under ASI irrespective of their employment size. Certain servicing units and activities like water supply, cold storage, repairing of motor vehicles and other consumer durables like watches etc. are covered under the Survey. Though servicing industries like motion picture production, personal services like laundry services, job dyeing, etc. are covered under the Survey but data are not tabulated, as these industries do not fall under the scope of industrial sector defined by the United Nations. Data Types:

Catalist

Title: Catalist

Catalist contains data from more than 240 million unique voting-age individuals from all 50 states and the District of Columbia. Data is compiled from a number of sources, including voter registration and Census data. Catalist provides two access points: the Q Tool and the M Tool. The Q Tool is a query to conduct extracts of the database, and the M Tool is used for uploading the researcher's data to the database and matching voter records from the researcher's dataset to the database information.

IMPORTANT: Use is limited to the UCB community only. All UCB users must apply for access via the D-Lab through the linked form. Staff at the D-Lab will contact the user when access is approved.

Data Use Agreement: < Apply for Access Here >

Nielsen

Conclusions

Acquiring and servicing data is like working with special collections. The material is expensive and requires additional effort to acquire, host, describe and service.

Data selectors need a knowledge of licensing, negotiations, file structures, and RDM.

Support from licensing, metadata specialists and systems administrators is essential.

Investing in a robust ticketing system potentially solves workflow issues.

The Wherefores

So undergraduates can write theses.

So grad students can write dissertations.

So faculty can write books and articles.

So librarians can support research teaching and learning

Making the World a Better Place

References

Church, Jim. "International survey data: Challenges and strategies for collection development." *DttP: A Quarterly Journal of Government Information Practice & Perspective*, 36 (1) (2008), p. 12-16.

Sheehan, Beth. "Assessing a Patron-Driven, Library-Funded Data Purchase Program." T*he Journal of academic librarianship.* 43 (1), (2017), p 49-54.

Smith, Kelly. "Problem Solved! Managing Electronic Resource Workflows Using Ticketing System Software," *Serials Review*, V42:1, (2016), p. 59-64.

Image Credits

1. Slide 1 - "Pit." https://en.wiktionary.org/wiki/pit
2. Slide 8 - "Data Purchase Program." https://www.library.illinois.edu/sc/purchase/
3. Slide 19 - "Breach of Agreement" http://picpedia.org/highway-signs/b/breach-of-agreement.html
4. Slide 23 - "Baby Sledding" http://www.afv.com/
5. Slide 42 - Ashwini shankar. "Make the world smile by being conscious of your waste" https://i.pinimg.com/originals/65/81/11/6581111d1ad676dd8aa128c26a2d1578.jpg