eDesiderata Forum: Licensing "Big Data" Summary Report

On November 16, 2016, the Center for Research Libraries held a virtual forum on the topic of licensing very large databases and datasets. The forum focused on commercial and open access data in three broad categories:

1. Business and financial data
2. Public opinion and population data
3. Geospatial data

The purpose of the event was to provide a better understanding of the challenges libraries face in licensing and acquiring such resources and a sense of how libraries can more effectively mediate between researchers and vendors/sources. The discussions explored how CRL might better support member library efforts to secure access to very large data resources for local researchers.

CRL member representatives will discuss possible steps CRL can take to address these concerns at the 2017 Global Resources Collections Forum, a webcast to be held in conjunction with CRL's Council of Voting Members meeting on April 21, 2017.

Session I: Business and Financial Data

Moderator: Cynthia Cronin-Kardon, Business Reference and Resource Development Librarian, Lippincott Library, Wharton School, University of Pennsylvania.

Panelists:

- Hilary Craiglow, Director, Walker Management Library, Vanderbilt University
- Alex Caracuzzo, Collections and Data Management Librarian, Harvard Business School Baker Library
- Barbara Esty, Senior Information Research Specialist, Harvard Business School

In the realm of business and financial data the landscape of resources incudes both off-the-shelf databases, like Bloomberg, WRDS, and others, and unique business and financial datasets obtained on an ad hoc basis from sources like Standard & Poor's and LexisNexis. In some instances a library will acquire the raw data for an off-the-shelf product.

Discussion points:

- For the major vendors that provide the large subscription data resources commonly used in economic, business and financial research (Bloomberg, Thomson Reuters, Morningstar, Standard & Poor's, and others) libraries are not the main clients. Therefore pricing can be quite high, and there is a tendency to purge historical data from their products.
- Commercial products often come with limitations, such as per-seat pricing, restriction of use to narrowly defined populations, and limits on downloads.
- Libraries are often unable to comply with certain vendor boilerplate licensing terms that are acceptable to non-academic customers. Terms that must be specially tailored to academic uses include, among others,

provisions governing vendor use of information on user identities and behaviors; local storage of the licensed data; and rights for long-term use and retention of the database.

- To keep the costs of major business databases manageable Vanderbilt has been able to form partnerships with non-traditional user groups on campus, like the development office.
- Work in the business and finance data licensing "ecosystem" often falls outside routine library eResource practices. Harvard librarians occasionally work on behalf of individual faculty to acquire specialized data.
- In some instances the data will be included in an off-the-shelf commercial product, but is needed in a format specifically tailored to the researcher's needs. In other instances it will be historical data that has been purged by the vendor from its commercial product. Historical pension fund data from Standard & Poor's *Money Market Directories* was one dataset mentioned as an example.
- Here the relationship with vendors is critical to success, and flexibility and creativity on the part of the library are necessary. Acquisition of such datasets is often on terms tailored to the intended research project, and data must often be customized to support a specific use.
- Therefore it is critical to know the scope and ultimate goals of the users' projects. Those goals will determine how to address matters like required format, but also duration of the period of access and rights for local hosting, retention and sharing of data.
- Librarians are likely to be more effective than researchers in codifying such terms, and in negotiating rights that satisfy user requirements. Researchers are usually focused on what data they need, and may often even know potential sources of that data, but still look to librarians to evaluate those sources.
- A close working relationship between researcher and librarian can yield other important benefits. While librarians know what questions to ask the data source, the enthusiasm of faculty and graduate student researchers, particularly those at large research universities, can provide new insights on data uses. Access to such insights can be valuable to vendors.
- It must also be determined in advance what tools and applications will be used by researchers to access, display and manipulate the data, such as Stata data analysis and statistical software or the open source Dataverse data management application developed by Harvard's Institute for Quantitative Social Science (IQSS) and University Library.
- Obtaining data and metadata samples for evaluation in advance of licensing is often necessary. Often, large purchased datasets will require technical support for downloading to local servers and other provisions for storage.
- The "support system" for licensing large business datasets should include a range of players, including researchers, university legal support, library subject matter experts, ¸IT staff.
- Dealing on a consortium basis with vendors of business and financial data might help solve some of the problems librarians face in licensing both off-the-shelf databases and unique datasets. Ideas shared included: educating vendors about the general needs and practices of academic users; producing and sharing boilerplate language for licenses for databases and datasets; and collective bargaining to negotiate pricing and terms for major, commonly used databases, scaled to the means of small and large institutions.

Session II: Public Opinion and Population Data

Moderator: Annelise Sklar, Social Sciences Collections Coordinator, University of California, San Diego Library

Panelists:

- Catherine Morse, Government Information, Law and Political Science Librarian, Stephen S. Clark Library, University of Michigan
- Lara Cleveland, Co-principal Investigator for Integrated Public Use Microdata Series-International, and Project Manager for the Minnesota Population Center, University of Minnesota
- Karen Hogenboom, Associate Professor of Library Administration, Scholarly Commons Librarian and Head of Scholarly Commons, University of Illinois at Urbana-Champaign

Session II focused on very large databases and datasets in two broad areas: demographic data, or data about people, and survey data on people's opinions, attitudes, and behaviors. Data types, research needs, resources, and challenges vary. As in the realm of business and finance, this domain includes open access data, off-the-shelf databases, and raw and customized data produced by organizations like Gallup and the Pew Research Center.

Discussion points:

- It is now common for population data providers to require a client library to individually identify all potential users and/or research projects, and to execute agreements defining permitted uses with each researcher. This is labor-intensive, and can violate longstanding library privacy policies.
- Problems also exist with data products themselves: documentation and metadata accompanying the data is often incomplete. And datasets acquired are often flawed by inconsistent, out-of-date or otherwise insufficiently contextualized values and variables.
- Population data are in demand by both well-established researchers and graduate students new to using very large data resources. Many researchers assume that all government-produced population data are available free of charge. Others are surprised that open data are often in a form less suitable to the intended use than the same data in subscription-based products.
- When public data are repackaged and sold by aggregators (East Views's *LandScan,* Reinvestment Fund's *PolicyMap,* and Oxford University Press's *Social Explorer*) the resulting "curated" products often provide added value. Librarians must decide whether to invest in collecting, normalizing, and maintaining open data locally, or to depend upon commercial providers and hope that they have a plan for long-term preservation.
- Few libraries have the resources to help researchers navigate the process of accessing highly restricted data. And many libraries are ill-equipped to deal with restricted data that cannot become part of the library collection.
- International census and population data present special challenges: they are often difficult to locate and obtain, requiring labor- and resource-intensive dealings with foreign government statistical agencies. Data specialists at Integrated Public Use Microdata Series (IPUMS), an open access repository of population data from various public sources based at the University of Minnesota do the difficult work of collecting and disseminating demographic microdata from U.S. and international censuses and surveys, processing much of this data to provide the structure and uniformity necessary for research use.
- IPUMS people negotiate individual licenses with over eighty countries to make their microdata available, and now maintain data dating from 1960 to the present day. Such work involves forging and maintaining

relationships with government agencies, and executing formal memoranda of understanding for each limited license.

- In many instances, agencies are reluctant to disclose domestic data; their reluctance stems from concerns about confidentiality, exposing flawed survey methodologies, or deficiencies in the data. In some instances the data can be politically sensitive. Statistical agencies also often fear the loss of revenue potentially to be earned from commercial data aggregators.
- There is a growing sense that traditional polling/survey data are inaccurate or unreliable, a sense only reinforced by the performance of polls in the recent U.S. presidential election. Moreover, traditional polling organizations like Pew and the National Opinion Research Center are having less success in reaching subjects by telephone, and are finding subjects increasingly resistant to disclosing their views.
- As a result, researchers are turning to non-traditional sources of opinion and behavioral data, like Google, Twitter and Facebook, and to political polling organizations like i360. Obtaining data from these sources is expensive, and raises a new set of issues for librarians. For example, the Twitter API used to aggregate data does not clearly indicate the scope of a given dataset, making replication of a given analysis using such data difficult. Morse suggested the value of a larger community discussion on these issues.
- Social science researchers are also looking to new sources for population distribution data created by remote sensing technologies and aerial imaging from satellites and drones. (Oak Ridge National Laboratory's *LandScan Global*, a dataset that merges census and satellite data, is an off-the shelf example distributed by East View.)
- A cooperative effort is needed to support the gathering and curating of population data by academy-facing organizations like IPUMs. Such an effort could support the digitization of historical census and population data that is now available only in paper and micro format.
- Panelists agreed that given the trends in use of population and public opinion data and growing demand for analytics services, a larger community discussion is needed on commercially available tools and social media data.

Session III: Geospatial Data

Moderator: Bernard Reilly, President, Center for Research Libraries

Panelists:

- *Amber Leahey,* Data and Geospatial Librarian, Scholars Portal, Ontario Council of University Libraries
- *John Faundeen,* Archivist, U.S. Geological Survey (USGS), Earth Resources Observation and Science (EROS) Center
- *Julie Sweetkind-Singer,* Assistant Director of Geospatial and Cartographic Services, Head of Branner Earth Science Map Library & Map Collections, Stanford University

Geospatial data includes information in many forms that has explicit geospatial positioning tying it to a specific place or region on earth. Geographic information systems (GIS) also provide a framework for visualizing data of many types--such as data relating to public health, demographics, weather, and natural resources-- in terms of the geolocation of their occurrence.

Discussion points:

- This is a complex data type with which many collection librarians are unfamiliar. Complexities include the immense size (terabyte scale) of datasets, continual and periodic editioning and updating, real-time data streams, and derivative products.
- Vendors of geospatial data have a variety of delivery and access mechanisms, and are often unacquainted with the academic landscape.
- With the entry of commercial satellite operators in the 1990s the sources of earth imagery have multiplied. This is also an area where the national security has both subsidized much work and created limitations due to sensitivity.
- There are benefits to licensing large geospatial datasets at the consortium level, as the experience of the Ontario Council of University Libraries attests. OCUL oversees licensing of geospatial data from agencies and sources at all levels of Canadian government on behalf of Ontario universities. Among the benefits: permits sharing of specialized licensing expertise that is not available at many individual universities; avoids the ad-hoc approach and one-off licenses that often occur at the institutional level; and offers economic and operational incentive to providers to negotiate with and distribute data through a central body, reducing the time and energy spent on their end.
- The OCUL effort also has the advantage of robust shared infrastructure provided by the GeoPortal (http://geo.scholarsportal.info) for hosting and maintaining geospatial data, maintained by the University of Toronto Scholars Portal.
- There is a complex interplay of public and commercial domains in the field of geospatial data. USGS has had some success in bringing commercially produced earth imagery into the public domain.
    - The USGS agreement with SICORP, a commercial distributor of earth imagery from a French satellite company, originally required the USGS sell the licensed images and share the revenue with the distributor. Eventually the agreement was re-negotiated and while the images remain under copyright, they are now available from USGS free of charge.
    - Under terms established by the Land Remote Sensing Policy Act of 1992, the USGS has successfully made historical earth-facing satellite images produced by certain commercial U.S.-based satellite operators like the American Hi-Resolution Satellite Company (and later "purged" from the operators' commercial products) publicly available free of charge or copyright restrictions.
- Conversely, Google and Amazon Services harvest and aggregate data from the USGS archive and incorporate that data in their own applications and products.
- Commercially produced geospatial datasets are often lacking essential provenance information and other important metadata. It can be difficult to ascertain the quality and even ownership of the data based on vendor-supplied information. Therefore it is advisable to obtain samples of data in advance of licensing or purchase, and share those with people at the library who can verify the quality and of both the data and the metadata.
- Data to be acquired must also be evaluated with regard to compatibility with existing university technology.
- With such a wide range of producers there are a variety of rights regimes and conditions that librarians must navigate. Licenses must be vetted to consider what rights to the data the library will require in the future as well as in the near term, such as the right to preserve the acquired data in a local repository and geospatial data infrastructure.

- There is widespread reliance on proprietary software, notably Esri's *ArcGIS*, to display and manage geospatial data. Panelists encouraged the building of open access GIS repositories and the use of open source software such as *GeoBlacklight* wherever possible, although few institutions have the capability and resources necessary to support custom-built repositories and therefore must rely on commercial options.

Conclusions

At present the landscape of business, population, public opinion and geospatial data is truly a "Wild West", where libraries must navigate a thicket of complexities to acquire and license resources for local researchers. One-off arrangements made by faculty and other researchers directly with data producers and vendors are common. Such arrangements often sacrifice potential economies of scale, broader sharing of benefits and long-term accessibility for the expediency of one-time access. In this realm libraries working closely with local researchers, IT support, legal counsel, and others can help achieve a greater return on the university's investment in data.

Certain concerns surfaced in all three sessions. While some of the problems highlighted (questions of legal jurisdiction, limitations on long-term retention of licensed data, etc.) are common to licensing in general, others are more typical in the world of very large data resources:

- A tendency to decentralize acquisition of data sets on campus, through one-off and ad hoc arrangements with vendors
- The commercialization of public data and the lack of transparency about derivative products
- Merging of content with tools, analytics and services: data is often less useful, even useless, removed from the vendor's platform
- High costs of local maintenance of very large data sets
- Tendency of vendors to purge historical data
- Important historical content remaining locked in paper format
- Purchase and licensing terms unsuited for academic purposes.

The discussions suggested four ways to strengthen library effectiveness in this realm:

1. *Collective dealings with vendors of critical databases and datasets:* A consortium approach to licensing of databases, comparable to that undertaken by OCUL, could improve terms and relieve individual U.S. libraries of much of the burden of dealing with vendors. CRL might work with OCUL, NERL, CRKN and others to secure favorable terms of access for key off-the-shelf databases, particularly those that provide historical data. Collective dealings with vendors on a national or international scale might also: create incentives for vendors to retain rather than purge historical data; help educate vendors on academic licensing norms and requirements; and achieve consensus on basic metadata standards for acquired data.
2. *Independent evaluation and analysis of key large databases and datasets:* Collecting and disclosure of reliable information about vendor technical infrastructure, and about measures taken by the major providers to ensure longevity of their data, could provide much-needed transparency in the big data marketplace. There is also a need for deep and ongoing analysis and evaluation of open access databases, to inform library investment in those data resources.
3. *Greater library support for non-commercial providers of data:* Academy-based and other not-for-profit organizations like IPUMS and the U.S. Geological Survey are expanding access to important data for

academic researchers. Working with those organizations could be a cost-effective way to bring about transfer to the public domain of historical data purged from commercial repositories and sources, and to subsidize conversion of historical financial and statistical data from print to digital form.

4. *Support for individual library dealings with vendors and data sources:* Community-wide pooling of expertise on the subject of big data could produce guidelines for individual library negotiations with producers and vendors. Cooperative work in this area might create boilerplate for specialized licensing terms to supplement those in LIBLICENSE and other model licenses.

Report posted January 10, 2017