In this issue

4 Editor’s notes
Karsten Boye Rasmussen

6 Social Science Data Archives
Kristin R. Eschenfelder, Morgaine Gilchrist Scott, Kalpana Shankar, and Greg Downey

20 Understanding Academic Patrons’ Data Needs through Virtual Reference Transcripts
Margaret Smith, Jill Conte, and Samantha Guss

27 Digitization, Data Curation, and Human Rights Documents
Amy Barton, Paul J. Bracke and Ann Marie Clark

35 First Forays into Research Data Dissemination
San Cannon, and Deng Pan
Editor’s notes

Our World and all the Local Worlds

Welcome to the first issue of Volume 40 of the IASSIST Quarterly (IQ 40.1, 2016). We present four papers in this issue. The first paper presents data from our very own world, extracted from papers published in the IQ through four decades. What is published in the IQ is often limited in geographical scope and in this issue the other three papers present investigations and project research carried out at New York University, Purdue University, and the Federal Reserve System. However, the subject scope of the papers and the methods employed bring great diversity. And although the papers are local in origin they all have a strong focus for generalization in order to spread the information and experience.

We proudly present the paper that received the ‘best paper award’ at the IASSIST conference 2015. Great thanks are expressed to all the reviewers who took part in the evaluation! In the paper ‘Social Science Data Archives: A Historical Social Network Analysis’ the authors Kristin R. Eschenfelder (University of Wisconsin-Madison), Morgaine Gilchrist Scott, Kalpana Shankar, and Greg Downey are reporting on inter-organizational influence and collaboration among social science data archives through data of articles published in IASSIST Quarterly in 1976 to 2014. The paper demonstrates social network analysis (SNA) using a web of ‘nodes’ (people/authors/institutions) and ‘links’ (relationships between nodes). Several types of relationships are identified: influencing, collaborating, funding, and international. The dynamics are shown in detail by employing five year sections. I noticed that from a reluctant start the amount of relationships has grown significantly and archives have continuously grown better at bringing in ‘influence’ from other ‘nodes’. The paper contributes to the history of social science data archives and the shaping of a research discipline.

The paper ‘Understanding Academic Patrons’ Data Needs through Virtual Reference Transcripts: Preliminary Findings from New York University Libraries’ is authored by Margaret Smith and Jill Conte who are both librarians at New York University, and Samantha Guss, a librarian at University of Richmond who worked at New York University from 2009-14. The goal of their paper is ‘to contribute to the growing body of knowledge about how information needs are conceptualized and articulated, and how this knowledge can be used to improve data reference in an academic library setting’. This is carried out by analysis of chat transcripts of requests for census data at NYU. There is a high demand for the virtual services of the NYU Libraries and there are as many as 15,000 annual chat transactions. There has not been much qualitative research of users’ data needs, but here the authors exemplify the iterative nature of grounded theory with data collection and analysis processes inextricably entwined and also using a range of software tools like FileLocator Pro, TextCrawler, and Dedoose. Three years of chat reference transcripts were filtered down to 147 transcripts related to United States and international census data. The unique data provides several insights, shown in the paper. However, the authors are also aware of the limitations in the method as it did not include whether the patron or librarian considered the interaction successful. The conclusion is that there is a need for additional librarian training and improved research guides.

The third paper is also from a university. Amy Barton, Paul J. Bracke, Ann Marie Clark, all from Purdue University, collaborated on the paper ‘Digitization, Data Curation, and Human Rights Documents: Case Study of a Library-Researcher-Practitioner Collaboration’. The project concerns the digitization of Urgent Action Bulletins of Amnesty International from 1974 to 2007. The political science research centered on changes of transnational human rights advocacy and legal instrumentation, while the Libraries’ research related to data management, metadata, data lifecycle, etcetera. The specific research collaboration model developed was also generalized for future practitioner-librarian collaboration projects. The project is part of a recent tendency where academic libraries will improve engagement and combine activities between libraries and users and institutions. The project attempts to integrate two different lifecycle models thus serving both research and curatorial goals where the central question is: ‘can digitization processes be designed in a manner that feeds directly into analytical workflows of social science researchers, while still meeting the needs of the archive or library concerned with long-term stewardship of the digitized content?’. The project builds on data of Urgent Action Bulletins produced by Amnesty International for indication of how human rights concerns changed over time, and the threats in different countries at different periods, as well as combining library standards for digitization and digital collections with researcher-driven metadata and coding strategies. The data creation started with the scanning and creation of the optical character recognized (OCR) version of full text PDFs for text recognition and modeling in NVivo software. The project did succeed in developing shared standards. However, a fundamental challenge was experienced in the grant-driven timelines for both library and researcher. It seems to me that the expectation of parallel work was the challenge to the project. Things take time.

In the fourth paper we enter the case of the Federal Reserve System. San Cannon and Deng Pan, working at the Federal Reserve Bank in Kansas City and Chicago, created a pilot for an infrastructure and workflow support for making the publication of research data a regular part of the research lifecycle. This is reported in the paper ‘First Forays into Research Data Dissemination: A Tale from the Kansas City Fed’ More than 750 researchers across the system produce yearly about 1,000 journal articles, working papers, etcetera. The need for data to support the research has been recognized, and the institution is setting up a repository and defining a workflow to support data preservation and future dissemination. In early 2015 the internal Center for the Advancement of Research and Data in Economics (CADRE) was established with a mission to support, enhance, and advance data or computationally intensive research,
and preservation and dissemination were identified as important support functions for CADRE. The paper presents details and questions in the design such as types of collections, kind and size of data files, and demonstrates influence of testers and curators. The pilot also had to decide on the metadata fields to be used when data is submitted to the system. The complete setup including incorporated fields was enhanced through pilot testing and user feedback. The pilot is now being expanded to other Federal Reserve Banks.

Papers for the IASSIST Quarterly are always very welcome. We welcome input from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution. We permit authors ‘deep links’ into the IQ as well as deposition of the paper in your local repository. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout: http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
June 2016
Editor
Abstract
This paper reports findings about inter-organizational influence and collaboration relationships among social science data archives over time, focusing on activities of institutions affiliated with the journal International Association of Social Science Information Services and Technology Quarterly (IASSIST Quarterly). We examine how archives interacted from 1976-2014 by tracing relationships described in articles published in IASSIST Quarterly.

Introduction
This paper reports findings about inter-organizational influence and collaboration relationships among social science data archives over time, focusing on activities of institutions affiliated with the journal International Association of Social Science Information Services and Technology Quarterly (IASSIST Quarterly). We examine how archives interacted from 1976-2014 by tracing relationships described in articles published in IASSIST Quarterly.

Keywords
Social science data archives, history, IASSIST Quarterly, social network analysis, collaboration

Introduction
Research disciplines increasingly rely on data archives and repositories to share results, advance their work, and support large-scale collaboration. While there have been numerous studies that examine the technologies and practices of how research fields and researchers develop and use data archiving and archives, there has been less attention paid to how data archives themselves as information institutions have adapted over time to evolving research trends, institutional changes, and funding models. Social Science Data Archives (SSDA) are exemplars of long-lived information infrastructures (broadly defined as the computing and technological resources and their supporting institutions that are designed to advance scientific inquiry) that have successfully adapted to such changes (Heim, 1980; O’Neill Adams, 2006). Understanding inter-organizational relationships among SSDA, funders and partner institutions over time is essential to understanding how SSDA have evolved to serve their user communities.

In this paper, we examine how such relationships have evolved, focusing on activities of institutions that appear in articles published in IASSIST Quarterly (IQ) from 1976-2014. Our larger goal is to better understand how SSDA have cooperated and competed to achieve their goals and draw out lessons learned that can be applied to the development and maintenance of contemporary cyberinfrastructures for research in the social sciences. We hope that lessons learned by SSDA may be useful to similar infrastructure projects in other fields.

This paper explores the following research questions:

1. Which institutions are most influential as depicted in IQ articles?
2. Which institutions collaborate the most as depicted in IQ articles?
3. To what degree are international relationships represented in IQ articles?
4. To what extent are highly collaborative institutions collaborating with each other?

Methods
Social network analysis (SNA) is a data collection and analysis approach useful for examining patterns in connections between people or social institutions. The results of SNA analysis are often depicted as a web of nodes (people or institutions) and relationships or links which represent the links between nodes. As Hansen, Schneiderman and Smith (2011) describe, SNA analysis may trace:

a. The number of unique links connected to a node. Nodes that have more links connected to other nodes may be more important or influential.
b. Changes in the patterns of connections between nodes.
  c. Variation in the types of links or between nodes.

To answer our research questions, we use SNA to analyze links between organizations or institutions (the nodes or our social network) as represented in IASSIST Quarterly (IQ) articles from 1976 through the end of 2014. We obtained back-issues of IQ from the IASSIST website, from archive.org or from the University of Wisconsin-Madison library. We did not include papers from the annual IASSIST conference proceedings because we only had access to the full text of conference presentations after 2000 (http://www.iassistdata.org/conferences).

We first identified all articles by issue according to the volume number and date listed on the bottom of the article. For each paper we identified the following nodes in each paper:

- Institutional home of author or co-author of IQ article,
- Institutions mentioned in the paper as influencers or collaborators (more on this below) by IQ authors
- Funders of projects described by IQ authors.

We identified the following major types of relationships, or links, as explicitly described in the papers:

1 Influencing relationship: When one node mentioned another node as being influential, including co-authorship when the authors were at different institutions. Social network analysis typically refers to nodes that have more attached links as having a higher degree of influence because a link between institutions represent opportunities for influence or evidence of influence. We examined the IASSIST networks to see which institutions were the most connected to other institutions. Arguably these linkages represent the potential for one institution to influence the practice of the other institution through sharing of knowledge or resources.

2 Collaborating institution relationship: When one node described a collaboration with another node. (see below)

3 Funding relationship: When a node provided funding (explicitly stated in the article). Funding relationships are shown in green on network graphs.

4 International relationship: We specifically noted data provider or collaborating institution relationships that crossed national boundaries. Because the Consortium of European Social Science Data Archives (CESSDA) is a pan-national organization, any relationship with CESSDA was marked as international.

We tracked influencing, collaborating, funding and international relationships among nodes in IQ using the tool NodeXL, an open source template for conducting SNA with data from Microsoft Excel (NodeXL 2015). We also tracked change over time of all of the four types of relationships. In order to show change over time, it is a common practice in longitudinal social network analysis papers to analyze data in multiyear sections rather than year by year. Data within a year often do not provide sufficient nodes and links to show a network of relationships. At the same time, analyzing the data as one large set (1976-2014) cannot show change. We determined that five year sections were sufficient to show both network relationships and change over time.

Operationalizing Influence and Collaboration

We coded the IQ articles in two ways: first broadly for influence, and then more narrowly for collaboration. Collaboration relationships are a narrower subset of influence relationships.

Influence: First, we coded broadly for relationships indicating influence among data archives. For the purpose of this project influence between nodes was included by one node's reference to another node as:

- an inspiration or model,
- part of a larger grouping of affiliated organizations,
- part of a collaboration,
- a provider of data,
- a provider of resources (staff, software), or
- a co-author on an IQ paper from a different institution

As we describe below, as we depicted in our (Eschenfelder et al., 2015) IASSIST 2015 conference paper, the influence analysis resulted in a large loose network of relationships in the IASSIST community. We argue that influence patterns show which data archives other archives talk about, or which data archives are most influential in the larger community.

Collaboration: In this narrower analysis, we examined a subset of influence relationships describing collaboration between data archives. To do so, we first developed a definition of collaboration and coding rules by conducting comparative analysis of three random samples of approximately ten IQ articles. From the analysis, the team identified the following types of collaboration relationships:

- New Collection/Service: Institution Y gets data from institution X when institution Y creates a new data collection or service. Or X allows Y to make X’s data accessible through a portal.
- New Entity/Project: Institutions X and Y create a new entity ‘project Z’ related to data.
- Cross National Surveys: Institution X describes participating in a cross national survey. Or a cross national survey project describes getting participation from nations X, Y and Z.
- Software/Processing: Institution X gets software from institution Y and creates a new data product or service. Or, institution X uses institution Y’s processing power to manipulate or manage data to provide a service.
• Collaborate to Collect Data: Institutions Y and Y collaborate to obtain data from individuals (PIs’ projects or from research subjects themselves).
• Research about Practice: Institution Y collects data from other data archives in order to compare practices and reports results.
• Mergers: Reports on acquisition of another data archive or its collection. Taking over a previous archive.
• Learning Materials: Institutions X, Y and Z collaborate to create learning materials.

In order for the description of an influence relationship in IQ to count as a collaboration, the depiction also needed to meet significance criteria. To count as significant, a description of a collaboration had to: have a header, have its own paragraph, be visually separated from other text, be mentioned multiple times within the article, or have at least 4 lines of text devoted to it.

IQ Articles included many descriptions of influence relationships that we did not count as collaboration including: X and Y co-authoring the IQ paper; providing curated links to data that lives at other institutions; compiling, indexing and coding data created through surveys run by others; simple descriptions of data deposit; descriptions of institution X getting data from institution Y and X writing a paper; mere descriptions of data available at institution X (even if the data is from other institutions); collaboration of subunits within the same larger organization; aspirational or planned collaborative activities; X describing how they contract out services to Y; and descriptions of teaching activities.

Coding Data for Influence and Collaboration
The research team first coded the relevant articles to identify influence relationships. A team member read through each article included in IQ, identified the names of nodes, and entered information about co-author, influence and funder relationships between nodes into an Excel template. Any given node typically had several links (influence relationships) and each link was listed in a separate row.

To code articles for the collaboration code, three members of the research team co-read and co-coded five separate samples of IQ influence articles from across different time periods in order to ensure we had agreement on the application of the coding rules. One team member then coded the remainder of the IQ articles and entered information about collaborations into a new Excel template. Each collaboration relationship was listed in a separate row.

Institutions and Relationships: 1976-2014
Figure 1 and Table 1 summarize data on the types of relationships reported in IQ articles from 1976-2014 including influence, collaborator, funder and international relationships. The number of nodes, or SSDA appearing in relationships grew over time from 12 in the 1976-1980 time period to 111 in the 2011-2014 period. The number of all types of relationships also grew over time, but not steadily. In particular, the number of funder relationships has been very up and down over the years. Moreover, the number of collaborator relationships reported in IQ grew steadily, but then dropped in the 2006-2010 reporting period.

Most nodes (or SSDA) in the data have a very small number of relationships. A small number of SSDA have a higher number of relationships. In this paper we focus on those SSDA with the higher number of relationships.
Which SSDA has had the most influence?
We examined the data for which SSDA had the most influence. Influence links included when one node referred to another node in an IQ article as an influence or inspiration, as part of a larger grouping of affiliated organizations, as part of a collaboration, as a provider of data, or a provider of resources (staff, software). One problem is that many influence relationships in our data stem from large projects with many partners. We call these instances ‘projects’. These projects are depicted in only one or two IQ articles, but include a high number of relationships.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Nodes</th>
<th># Influence Links</th>
<th># Collaborator Links</th>
<th># Funder Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall 1976-2014</td>
<td>653</td>
<td>285</td>
<td>196</td>
<td>145</td>
</tr>
<tr>
<td>1976-1980</td>
<td>12</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1981-1985</td>
<td>37</td>
<td>27</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>1986-1990</td>
<td>81</td>
<td>69</td>
<td>26</td>
<td>12</td>
</tr>
<tr>
<td>1991-1995</td>
<td>76</td>
<td>54</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>1996-2000</td>
<td>108</td>
<td>77</td>
<td>37</td>
<td>25</td>
</tr>
<tr>
<td>2001-2005</td>
<td>130</td>
<td>119</td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td>2006-2010</td>
<td>98</td>
<td>81</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>2011-2014</td>
<td>111</td>
<td>130</td>
<td>31</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 1: Relationships over time

We created a measure called ‘influence ratio’ to depict those institutions with the most influence. This measure helped us identify institutions with many influence relationships over time, in addition to those with many influence relationships. The influence ratio captures this measure of an institution’s influence across different articles. The ratio divides the number of total articles in which institution X appears by the total number of influence links between institution X and others. Table 2 reports the institutions with the

<table>
<thead>
<tr>
<th>Overall 1976-2014: SSDA With The Highest Number Of Influence Links To Other SSDA</th>
<th># of Influence Links</th>
<th># Articles in Which the Institution is Coded</th>
<th>Influence Ratio: Articles/Links</th>
<th>Links/articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-University Consortium for Political and Social Research (ICPSR)</td>
<td>30</td>
<td>22</td>
<td>0.73</td>
<td>1.36</td>
</tr>
<tr>
<td>Zentralarchiv fur Empirische Sozialforschung (ZA)</td>
<td>16</td>
<td>9</td>
<td>0.56</td>
<td>1.78</td>
</tr>
<tr>
<td>UKDA</td>
<td>32</td>
<td>17</td>
<td>0.53</td>
<td>1.88</td>
</tr>
<tr>
<td>University of Edinburgh</td>
<td>13</td>
<td>6</td>
<td>0.46</td>
<td>2.17</td>
</tr>
<tr>
<td>International Social Survey Program (ISSP)</td>
<td>21</td>
<td>7</td>
<td>0.33</td>
<td>3</td>
</tr>
<tr>
<td>US Federal Reserve</td>
<td>14</td>
<td>4</td>
<td>0.29</td>
<td>3.5</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>15</td>
<td>4</td>
<td>0.27</td>
<td>3.75</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>19</td>
<td>2</td>
<td>0.11</td>
<td>9.5</td>
</tr>
<tr>
<td>The Pew Forum</td>
<td>19</td>
<td>2</td>
<td>0.11</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Table 2: 1976-2014 Nodes with Most Influence Links
highest influence ratio from 1976-2004. Large data archives appear at the top with ICPSR leading with a ratio of .73, followed by the ZA and UKDA with ratios of .56 and .53 respectively.

Which SSDA has collaborated the most?
We examined the data to see which SSDA has the most collaborative relationships. Collaborative relationships were a narrower subset of influence relationships that involved a specific set of relationship types including: creating new collections or services, new projects, cross national surveys, sharing of software or processing, new data collections, comparisons of practices or mergers.

Because we were interested in reporting on institutions with many collaborative relationships over time, we focus our analysis on those institutions depicted in more than one IQ article. Table 3 lists the institutions with the most collaborative relationships in terms of the same article/links ratio used above. Again, large data archives appear at the top of the list with ICPSR having a collaboration ratio of 1 and UKDA having a ratio of .85. The US Census Bureau had the next highest ratio with .43.

<table>
<thead>
<tr>
<th>Overall 1976-2014: Institution with the Highest Number Of Links To Other Nodes (c)</th>
<th># of Collaboration Links (Coding 2)</th>
<th>Number of IQ articles in which institution is mentioned</th>
<th>Collaboration Ratio (articles/ links)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICPSR</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>UKDA</td>
<td>13</td>
<td>11</td>
<td>.85</td>
</tr>
<tr>
<td>US Census Bureau</td>
<td>7</td>
<td>3</td>
<td>.43</td>
</tr>
<tr>
<td>International Social Survey Program (ISSP)</td>
<td>7</td>
<td>2</td>
<td>.29</td>
</tr>
<tr>
<td>Integrated Library and Survey-data Extraction Service (ILSES)</td>
<td>11</td>
<td>2</td>
<td>.19</td>
</tr>
<tr>
<td>East Asian Business and Development (EABAD) archive</td>
<td>7</td>
<td>2</td>
<td>.29</td>
</tr>
</tbody>
</table>

Table 3: Institutions with the Most Collaborating Relationships across Multiple IQ Articles 1976-2014

Who has funded whom?
We examined funding relationships and nodes to determine which funders had the most relationships from 1976-2014. Table 4 below shows the funding nodes with the most relationships in the 1976-2014 period. JISC was the top funding node with 18 reported relationships, followed by NSF and ESCR with 12 relationships each.

<table>
<thead>
<tr>
<th>Funder</th>
<th>Number of Funding Links 1976-2014</th>
<th>Number of IQ articles in which funder is mentioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Information Systems Committee (JISC)</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>National Science Foundation</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Economic and Social Research Council (ESRC)</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Higher Education Support Project (HESP)</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>National Institute on Aging</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Library of California</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>National Institute of Child Health and Human Development</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>California Department of Finance</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4 shows several instances of projects with many funding links that are only reported in one article (e.g., Library of California). These instances represent IQ articles that describe projects in which a funder funded a project with multiple partner nodes. In the next section, we continue by providing more detailed data on those nodes with the most influence and collaboration relationships during each of eight five year time periods of our study. We also report on international relationships during each period.

**Period 1: 1976-1980**

The early time period of 1976-1980 saw the lowest number of nodes (N=12) and the lowest number of influence and collaboration links (N=13, N=5). At this stage in IQ's history, most articles tended to simply describe activities at an author's institution, and few described cooperative activities.

**Influence**: Roper had the highest number of influence relationships (6), followed by University of Iowa (4) and then Yale, Williams College and the University of Connecticut (all of which were affiliated with Roper Center – 3 each).

**Collaboration**: In the narrower collaboration measure, only University of Iowa and the Roper Center showed more than one collaboration link during this period.

### Table 5: 1976-1980 Number and Type of Relationships

<table>
<thead>
<tr>
<th>Most Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roper (6)</td>
<td>(2 each) Roper Center;</td>
<td></td>
</tr>
<tr>
<td>University of Iowa (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yale University; Williams College; University of Connecticut (3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**International**: IQ articles did not depict any international collaborations during the 1976-1989 period.

**Period 2: 1981-1985**

The second time period saw a growth in nodes (N=37) and influence and collaboration relationships (N=27, N=10). This period also saw a rise in international relationships.

**Influence**: In this period, ICPSR and the US Census Bureau had the highest number of influence links (6) followed by Rand Corporation (4).

**Collaborations**: IQ articles still reported very few collaboration relationships. Four institutions reported two collaborations each: Center for Human Resource Research, Indian Council of Social Science Research (ICSSR), the International Federation of Data Organizations for the Social Sciences (IFDO), Norwegian Social Science Data Services, Zentralarchiv fur empirische Sozialforschung (ZA).

### Table 6: 1981-1985 Number and Type of Relationships

<table>
<thead>
<tr>
<th>Top Most Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICPSR; US Census Bureau (6 each)</td>
<td>(2 each) ICSSR, IFDO, NSD, ZA</td>
<td></td>
</tr>
<tr>
<td>2nd Most Relationships</td>
<td>Rand (4)</td>
<td></td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>(all with 3) (NORC); NSD, NARA; US DoE; Bonneville</td>
<td></td>
</tr>
</tbody>
</table>

**International Collaborations**: This period saw three international collaborations. Both involved partners in Europe or partnerships with international organizations such as IFDO.
Table 7: 1981-1985 International Collaborations

<table>
<thead>
<tr>
<th>Year</th>
<th>Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td>Norwegian Social Science Data Service and European Consortium on Political Research (ECPR)</td>
</tr>
<tr>
<td>1985</td>
<td>Two projects between Zentralarchiv fur empirische Sozialforschung and the International Federation of Data Organizations for the Social Sciences (IFDO)</td>
</tr>
</tbody>
</table>

Period 3: 1986-1990

The third period showed growing interactivity among SSDA. Nodes increased to 81. Influence relationships increased to 69 and collaboration relationships grew to 26.

Influence: The International Social Survey Programme (ISSP) had the highest number of influence links (13) followed by Australian National University (6).

Collaboration: The East Asian Business and Development (EABAD) archive had the most collaboration links (5) stemming from a large multi-institution project. The US Census had 4 reported collaborations in this period.

<table>
<thead>
<tr>
<th>1986-1990: Institutions with the Highest Number of Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Most Relationships</td>
<td>International Social Survey Program (ISSP) (13)</td>
<td>East Asian Business and Development (EABAD) (5)</td>
</tr>
<tr>
<td>2nd Most Relationships</td>
<td>Australian National University (6)</td>
<td>US Census Bureau (4)</td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>(5 each)National Opinion Research Center (NORC); East Asian Business and Development Research Archive (EABAD)</td>
<td>(3 each) TARKI (Hungarian Social Science Information Center)</td>
</tr>
<tr>
<td>4th Most Relationships</td>
<td>(4 each) ICPSR; University of Amsterdam; University of Alberta; Office of Population Census and Surveys UK; Hunter College CUNY; University of Mannheim</td>
<td></td>
</tr>
</tbody>
</table>

International:

This period’s IQ articles described an international collaboration between CELADE in Chile and the ICRC in Canada in 1989. In 1990 the East Asian Business and Development (EABAD) Archive at UC Davis reported on a collaborative project involving numerous partners in Asian nations. The ECPR, an international scholarly political science association located in Essex UK reported collaborations with two European universities.

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>United Nations Latin American Demographic Center(CELADE) in Chile and International Development Research Center of Canada (ICRC)</td>
</tr>
</tbody>
</table>
1990

East Asian Business and Development (EABAD) Archive at UC Davis and a series of partners including University of Hong Kong, National University of Singapore, Tunghai University in Taiwan and the China Credit Information Service
ECPR (UK) and both University of Mannheim and University of Amsterdam

Table 9: 1986-1990 International Collaborations

**Period 4: 1991-1995**

The fourth time period saw a decline in reported activity among SSDA. The number of nodes fell from 81 to 76, the number of influence links fell from 69 to 54, and the number of collaborations fell from 26 to 21.

**Influence:** University of Manchester and the Manchester Computing Center had the highest number of links in this period (10).

**Collaboration:** University of California Davis had the most collaboration links in IQ (5).

<table>
<thead>
<tr>
<th>1991-1995: Institutions with the Highest Number of Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Most Relationships</td>
<td>University of Manchester and Manchester Computing Center (10)</td>
<td>UC Davis (5)</td>
</tr>
<tr>
<td>2nd Most Relationships</td>
<td>University of Wisconsin-Madison (8); (2 each) Bringham University SfNY, East Asian Business and Development (EABAD); ICPSR; Lehman College; ROADS; US Census Bureau</td>
<td></td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>University of Illinois Urbana Campaign (7)</td>
<td></td>
</tr>
<tr>
<td>4th Most Relationships</td>
<td>(5 each) US Census Bureau; University of California Davis</td>
<td></td>
</tr>
<tr>
<td>5th Most Relationships</td>
<td>University of Missouri St Louis (4)</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: 1991-1995 Number and Type of Relationships

**International:** This period’s IQ articles described only one international collaboration between the University of Ulster and the United Nations University (a think tank and post graduate education institution associated with the United Nations and located in Japan).

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>University of Ulster and the United Nations University (Japan)</td>
</tr>
</tbody>
</table>

Table 11: 1991-1995 International Collaborations

**Period 5: 1996-2000**

Activity grew again during the 1996 to 2000 time period. Nodes grew to 108, influence links grew to 77 and collaborations grew to 37.

**Influence:** In this period, the UKDA had 11 influence links and the ZA had 7 links.

**Collaborations:** The Integrated Library and Survey-data Extraction Service (ILSES) appeared as the largest collaborator during this period with eleven links. The Data Liberation Initiative followed with seven links.
1996-2000: Institutions with the Highest Number of Relationships

<table>
<thead>
<tr>
<th>Top Most Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKDA (11) ILSES (11)</td>
<td>Integrated Library and Survey-data Extraction Service (ILSES) (11)</td>
<td></td>
</tr>
<tr>
<td>2nd Most Relationships</td>
<td>Zentralarchiv fur Empirische Sozialforschung (ZA) (7)</td>
<td>Data Liberation Initiative (7)</td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>University of Minnesota; Conference of Rectors and Principals of Quebec Universities</td>
<td>ISSP (5)</td>
</tr>
<tr>
<td>4th Most Relationships</td>
<td>University of Quebec; Sociometrics Corporation</td>
<td>Royal Statistical Society (4)</td>
</tr>
</tbody>
</table>

Table 12: 1996-2000 Number and Type of Relationships

International: International collaborative relationships described in this period include three descriptions of Integrated Library and Survey-data Extraction Service projects with multiple European partners in 1997, 1998 and 2000

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>Integrated Library and Survey-data Extraction Service (ILSES) and partners including ProGAMMA, Groningen, SWIDOC in Amsterdam, University of Amsterdam, Zentralarchiv fur Empirische Sozialforschung (ZA), Trinity College and CIDSP in Grenoble</td>
</tr>
<tr>
<td>1998</td>
<td>Integrated Library and Survey-data Extraction Service (ILSES) and partners including ProGAMMA, Groningen, Zentralarchiv fur Empirische Sozialforschung (ZA), University of Amsterdam, BSDP, Trinity College</td>
</tr>
<tr>
<td>2000</td>
<td>International Social Survey Program (ISSP) and partners including National Opinion Research Center (NORC), Zentrum fur Umfragen, Methoden und Analysen (ZUMA), Zentralarchiv fur Empirische Sozialforschung (ZA), National Center for Social Research, Research School of Social Sciences</td>
</tr>
</tbody>
</table>

Table 13: 1996-2000 International Collaborations

Period 6: 2001-2005
The sixth period saw growth in reported activity as nodes rose to 130, influence relationships rose to 119, and collaborations rose to 53.

Influence: Penn State and Pew had the highest number of influence links with 18 each.

Collaboration: The Association of Religious Data Archives had 20 reported collaborations in this period.

<table>
<thead>
<tr>
<th>2001-2005: Institutions with the Highest Number of Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Most Relationships</td>
<td>(18 each) Pennsylvania State University; the Pew Forum; also Association of Religion Data Archives (ARDA) (20)</td>
<td>Association of Religion Data Archives (ARDA) (20)</td>
</tr>
</tbody>
</table>
Table 14: 2001-2005 Number and Type of Relationships

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Sociological Data Archive (SDA) at the Institute of Sociology in Prague and nine other European collaborators. Institute for Quality of Life Research Bucharest (IQLR), UKDA and ZA.</td>
</tr>
<tr>
<td>2002</td>
<td>Collection of Census Data and Resources (UK) and the Fraunhofer Institut Autonome Intelligente Systeme (Germany)</td>
</tr>
<tr>
<td>2003</td>
<td>Finish Social Science Data Archive (Finland) and ICPSR (USA)</td>
</tr>
<tr>
<td>2005</td>
<td>Association of Religion Data Archives (USA) and 20 other partners, mostly based in US but some European</td>
</tr>
</tbody>
</table>

Table 15: 2001-2005 International Collaborations

**Period 7: 2006-2010**

The seventh time period was another period of decline. Articles only mentioned 98 nodes, and 81 influence links. And we saw a large decline in collaboration links described from 53 in the earlier period to merely 13 in this period.

**Influence:** The Finnish Social Science Data Archive and the University of Ljubjana showed the most influence links with 10 and 9 links respectively.

**Collaboration:** University of Edinburgh, University of Oxford, University of South Hampton and London School of Economics all showed three collaborations each.

<table>
<thead>
<tr>
<th>2006-2010: Institutions with the Highest Number of Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Most Relationships</td>
<td>Finnish Social Science Data Archive (10)</td>
<td>(3 each) University of Edinburgh; University of Oxford; University of South Hampton; London School of Economics</td>
</tr>
<tr>
<td>2nd Most Relationships</td>
<td>University of Ljubjana (9)</td>
<td></td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>University of Minnesota (7)</td>
<td></td>
</tr>
<tr>
<td>4th Most Relationships</td>
<td>(6 each) University of Windsor; University of Albany; University of Cape Town</td>
<td></td>
</tr>
</tbody>
</table>
Table 16: 2006-2010 Number and Type of Relationships

| 5th Most Relationships | (5 each) UKDA; US Federal Reserve; MIT |

Table 17: 2006-2010 International Collaborations

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>African Association of Statistical Data Archivists (AASDA) and the International Household Survey Network (IHSN)</td>
</tr>
</tbody>
</table>

**International:** The number of international collaborations described in IQ also dropped in 2006-2010. We found only one: between the African Association of Statistical Data Archivists and the International Household Survey Network in 2007.

Table 18: 2011-2014 Number and Type of Relationships

<table>
<thead>
<tr>
<th>2010-2014: Institutions with the Highest Number of Relationships</th>
<th>Influence</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Most Relationships (24 each) Lithuanian Center for Social Research; Vilnius University</td>
<td>Nestor Memorandum of Understanding (13)</td>
<td></td>
</tr>
<tr>
<td>2nd Most Relationships (11 each) Grottingen State University; Cologne University of Applied Sciences</td>
<td>(3 each) American National Election Study (ANES); Lithuanian Humanities and Social Science Data Archive (LiDA); VOICES Project/EQUALAN; Wienmer Institute for Social Science Data Documentation and Methods (WISDOM)</td>
<td></td>
</tr>
<tr>
<td>3rd Most Relationships</td>
<td>UKDA (9)</td>
<td></td>
</tr>
<tr>
<td>4th Most Relationships (8 each) US Federal Reserve; Wienmer Institute for Social Science Data Documentation and Methods (WISDOM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th Most Relationships</td>
<td>Council of European Social Science Data Archives (CESSDA) (7)</td>
<td></td>
</tr>
</tbody>
</table>

**Period 8: 2011-2014**

Growth increased in the final analysis period with 111 nodes, 130 influence relationships and 31 collaborations.

**Influence:** The Lithuanian Center for Social Research and Vilnius University scored highest in terms of number of influence links with 24 each.

**Collaboration:** The node with the highest number of collaborators was the Nestor Memorandum of Understanding (13). This stems from one large project with many European collaborators.

International: This period’s IQ articles described three major international collaborations. Two were largely European, but one stretched between Taiwan and Norway.
Table 19: 2011-2014 International Collaborations

<table>
<thead>
<tr>
<th>Year</th>
<th>International Collaboration Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>VOICES Project/EQUALAN and 3 European partners, Lithuanian Humanities and Social Science Data Archive (LIDA) and 2 European partners</td>
</tr>
<tr>
<td>2012</td>
<td>Center for Survey Research, Taiwan (SRDA) and the Norwegian Social Science Data Services Nestor Memorandum of Understanding among 13 European partners</td>
</tr>
</tbody>
</table>

Summary Discussion
In this section we answer the paper’s research questions by examining our data over time.

Research Question 1: Who is the most influential SSDA as depicted in IQ?

The most influential SSDA (according to our measure of influence) are ICPSR, the ZA, UKDA, and the University of Edinburgh. The ISSP project was also very influential but not at the level of the other four. In other words, within the IQ environment, these nodes were most often mentioned, across different articles, as having an influence via collaboration, data sharing, or just serving as an example for others.

<table>
<thead>
<tr>
<th>Overall 1976-2014: SSDA With The Highest Number Of Influence Links To Other SSDA</th>
<th>Influence Ratio: Articles/Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-University Consortium for Political and Social Research (ICPSR)</td>
<td>0.73</td>
</tr>
<tr>
<td>Zentralarchiv fur Empirische Sozialforschung (ZA)</td>
<td>0.56</td>
</tr>
<tr>
<td>UKDA</td>
<td>0.53</td>
</tr>
<tr>
<td>University of Edinburgh</td>
<td>0.46</td>
</tr>
<tr>
<td>International Social Survey Program (ISSP)</td>
<td>0.33</td>
</tr>
<tr>
<td>US Federal Reserve</td>
<td>0.29</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>0.27</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>0.11</td>
</tr>
<tr>
<td>The Pew Forum</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 20: Nodes with the Most Influence Relationships

Research Question 2: Who collaborates the most as depicted in IQ?

The SSDA with the most collaborative relationships according to our measure is ICPSR, with UKDA close behind it. The US Census Bureau also appears prominently. This means that as depicted in IQ articles, ICPSR and UKDA had the most relationships that met our criteria for collaboration described earlier.

<table>
<thead>
<tr>
<th>Overall 1976-2014: Institution with the Highest Number Of Links To Other Nodes (c)</th>
<th>Collaboration Ratio (articles/links)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICPSR</td>
<td>1</td>
</tr>
</tbody>
</table>
Research Question 3: To what degree are international collaborations represented in IQ articles?

As shown in Table 22, the number of international collaborative links varies widely across the time periods of the study. International collaborative relationships have grown over time but seen two periods of decline (1991-1995 and 2006-2010).

Table 22: Collaborations and International Collaborations Over Time

<table>
<thead>
<tr>
<th>Time Period</th>
<th># All Collaboration Links</th>
<th># International Collaboration Links</th>
<th>Ratio Intl/All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976-1980</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1981-1985</td>
<td>10</td>
<td>3</td>
<td>.3</td>
</tr>
<tr>
<td>1986-1990</td>
<td>26</td>
<td>8</td>
<td>.31</td>
</tr>
<tr>
<td>1991-1995</td>
<td>21</td>
<td>4</td>
<td>.19</td>
</tr>
<tr>
<td>1996-2000</td>
<td>37</td>
<td>16</td>
<td>.43</td>
</tr>
<tr>
<td>2001-2005</td>
<td>53</td>
<td>30</td>
<td>.57</td>
</tr>
<tr>
<td>2006-2010</td>
<td>13</td>
<td>1</td>
<td>.07</td>
</tr>
<tr>
<td>2011-2014</td>
<td>31</td>
<td>18</td>
<td>.58</td>
</tr>
</tbody>
</table>

To examine whether the proportion of international collaborative relationships to overall collaborative relationships has grown in IQ articles, we plotted a ratio of international collaborations over all collaborations by time period.

Collaboration downturns: The results shown in Figure 2 suggest that the proportion of international collaborations is growing over time, with periods of downturn (e.g., 1991-1995, 2006-2010). This pattern of growth and dips in growth parallel the overall pattern of collaboration activity shown in Figure 1 and Table 22 above. Other kinds of analysis might yield more insight into whether these collaboration downturns were due to financial pressures or other factors, for example as economies go into recession, or governments cut funding on education fewer resources may be available to support travel to collaborate with other SSDA. Alternatively, SSDA staff may be more under resourced and have less time to write up articles describing collaborations to IQ.

Research Question 4: To what extent are highly linked institutions interlinked with each other?

We took the institutions with the most relationships from Table 21 and looked to see to what degree they had collaboration relationships with other ‘most collaborative’ nodes versus collaborative relationships with other nodes. We found that the ‘most collaborative’ nodes did not collaborate with each other; rather, they collaborated with other nodes with lower collaboration scores. This suggests that highly connected nodes are mostly serving as hubs, linking outwards toward less connected nodes.

Conclusions

The data generated by this analysis provide one limited view of the relationships among SSDA and related institutions. There was partial overlap between most influential and most collaborative SSDA. Our analysis show that the most influential SSDA include ICPSR, ZA, UKDA and University of Edinburgh. The most collaborative SSDA were ICPSR and the UKDA followed by the US Census Bureau. Our analysis also found that the top ranked collaborative SSDA did not collaborate with each other but with other institutions. Top
IASSIST members would seek to share information about their projects with other members. Secondly, while the data set of relationships is incomplete, it is still representative of the most innovative and noteworthy relationships in the study period. Future research could enrich understanding of influence and collaboration relationships between SSDA by adding data from IASSIST annual meeting conference proceedings. At this time, only data from 2000 and later is available via the IASSIST website. It is much more difficult to draw conclusions about lack of collaboration or competition from this data set or reasons why relationships might have ended. To learn more about these issues, we plan to rely more on historical documentation from case studies of individual SSDA.

SSDA are complex institutions with a long history (Heim, 1980; O’Neill Adams, 2006). This IQ analysis provides one view of how SSDA have developed and maintained relationships with their funders, and each other over time to innovate, serve their user bases, and grow their products and services. This paper has described and summarized inter-organizational relationships among SSDA and funders from 1976-2014 as depicted in articles published in IASSIST Quarterly. As one component of a larger study about the history of social science data archives (SSDA) and the field of social science data archiving, the data provides an important broad view of relationships among data archives in the shaping of a research discipline. We would suggest that this study, even though it is partial in scope, represents the kinds of institutional analyses that could yield insight for other kinds of data repositories and institutions in how they are developing and leveraging their own institutional networks and the implications such networks have for growth and maintenance.

References

Acknowledgements
This project has been supported by the Alfred P. Sloan Foundation, the University of Wisconsin Alumni Research Foundation, University College Dublin Seed Funding Scheme, and the American Society for Information Science and Technology History Fund.

Notes
1. Corresponding author is Professor Kristin R. Eschenfelder, School of Library and Information Studies, University of Wisconsin-Madison, eschenfelder@wisc.edu
2. As explained by IQ editors, the year dates of publication for an article are not always precise because publishing is sometimes behind schedule. Analysis by five year periods may therefore be a better depiction of trends than analysis by specific year.
3. Where there were more than two authors, all authors’ institutions were listed as collaborating with one-another. When one author was associated with multiple institutions, each institution was listed as having all the relationships referenced in the article. However, they were not listed as collaborating with each other.
Understanding Academic Patrons’ Data Needs through Virtual Reference Transcripts:

Preliminary Findings from New York University Libraries

by Margaret Smith¹, Jill Conte², Samantha Guss³

Abstract

New York University (NYU) Libraries has an extremely high-volume chat reference service. This popularity presents a unique opportunity for gaining insight into library patrons’ conceptualizations of their data reference needs and how these needs are changing. Through analysis of three years’ worth of chat transcripts, we began to explore user needs and familiarity related to locating secondary data and statistics, performing data analysis, and using existing data services. Ultimately, we focused our analysis on requests for census data. This article discusses, in detail, the methods, preliminary results, limitations, and proposed next steps of our investigation. Our final goal is to contribute to the growing body of knowledge about how information needs are conceptualized and articulated, and how this knowledge can be used to improve data reference in an academic library setting.

Keywords: academic libraries, data reference, grounded theory, virtual reference services, chat transcripts

Introduction

NYU Libraries serves the NYU ‘Global Network University’, the main campus of which is situated in Greenwich Village, next to Washington Square Park, in Lower Manhattan. The NYU Polytechnic School of Engineering is housed nearby, in downtown Brooklyn, and NYU has portal campuses in Abu Dhabi and Shanghai, as well as 11 smaller global academic centers where students study away for a semester or year. NYU enrolls approximately 45,000 students (half of whom are undergraduate students), and employs approximately 3,000 teaching faculty. Bobst Library is the flagship of the NYU Libraries’ system, with 12 publicly accessible floors, 6 million volumes, and seating for 3,000.

The library’s urban location and proportionately small seating capacity, combined with the area’s above-average commute time and a user community spanning the globe, lead to high demand for NYU Libraries’ virtual library services. Our chat reference service is extremely busy; we receive approximately 15,000 chat transactions annually, 30-40 a day on average, mostly occurring between the hours of 9am and midnight, New York City local time. The average duration of a chat conversation is 16 minutes.

This popularity offers a unique opportunity for gaining insight into library users’ conceptualizations of their data needs and how these needs are changing. Through analyzing three years’ worth of chat reference transcripts, we began to explore user needs and familiarity related to locating secondary data and statistics, performing data analysis, and using existing data services, focusing on the way patrons initially ask data questions. While existing scholarship has addressed the theory and practice of data reference (Gerhan, 1999; Kellam and Peter, 2011), very little empirical research to date has qualitatively explored users’ articulations of their data needs (Wang, 2013). This project is unique in that it employs transcripts of actual reference transactions, as opposed to user statistics, performing data analysis, and using existing data services, focusing on the way patrons initially ask data questions. While existing scholarship has addressed the theory and practice of data reference (Gerhan, 1999; Kellam and Peter, 2011), very little empirical research to date has qualitatively explored users’ articulations of their data needs (Wang, 2013)

... patrons and librarians often use the word ‘data’ casually when discussing databases or information in general.
To analyze these transcripts, we used Dedoose, a web-based application developed to perform mixed-methods analyses in the social sciences. Dedoose allows us to categorize each transcript using controlled descriptors—for example, to indicate whether a transcript should be included or excluded from a sample, and also to apply qualitative codes to excerpts of text within the transcripts. These descriptors and codes could then be cross-tabulated, analyzed, and visualized in various ways. In combination, these tools—FileLocator Pro, TextCrawler, and Dedoose—were extremely effective for selecting a sample of transcripts, for protecting the privacy of individuals involved, and for classifying and analyzing the transcripts within a sample.

The process of gathering data-related reference transactions, however, was a non-trivial task. Even generating a starting search strategy required careful consideration of disambiguation. For example, we quickly realized that a search for the phrase, ‘number of’, would also retrieve results where a librarian or patron mentions the call number of a book. After a few minor tweaks to minimize these mismatches, our search strategy settled on this:

- data OR statistics OR stats OR GDP OR demographics OR census OR mortality OR GIS OR quantitative OR numeric OR SPSS OR Atlas OR Atlas OR NVivo OR qualitative OR vivo OR “Data Services” OR data.services@nyu.edu OR “Data Service Studio” OR “data.service@nyu.edu” OR STATA

This limited the number of transcripts substantially, but still retrieved an immense number of transcripts that were not data-related. For example, patrons and librarians often use the word ‘data’ casually when discussing databases or information in general. Additionally, there were quite a few hits where the patron was asking for help locating or accessing a book or article that had one or more of our search terms in its title, yet the resource itself was not data-related (e.g., a quantitative study related to nursing). There were also cases where the physical space of our Data Services department was referenced, but not in regard to data needs (e.g., complaints of an unruly patron or broken computer in that area).

In order to ensure that the sample contained as many data-related results as possible, we read through the transcripts, looking for actual relevance to data, and assigned an inclusion or exclusion descriptor to each one. Even so, we ended up with 950 data-related transcripts from just one year’s worth of transcripts. So we further refined our inclusion/exclusion criteria to omit those data-related transcripts involving ‘known item’ questions, such as a patron asking for help locating a specific financial report that contained data they had found via Google. While sometimes these patrons seemed clearly interested in the data that the report contained, it was often difficult to say whether this was definitively the case, or whether they were more interested in the report as a whole. We applied these new descriptors to the sample. At this point, 633 transcripts remained, a large proportion of which still involved questions about specific databases for business and financial information.

At a loss for ideas of other wide-sweeping exclusions we could make, we made a first pass at creating descriptors and codes for the transcripts in this sample. We read through them separately, coming up with lists of descriptors/codes that seemed potentially relevant, such as which specific resources were mentioned, the general subject area of the query, how accurate the librarian’s answer was (on a numeric scale), and how satisfied the patron seemed (on a numeric scale). We then discussed our experiences as a group and quickly realized the overwhelming effort that would be necessary to apply multiple, quantitative descriptors to a sample of this size. We decided to drop nearly all of the descriptors, and instead, apply codes within the text of each transcript, indicating the presence of different characteristics, like ‘inaccurate answer’ or ‘patron satisfaction.’ This was a speedier process, and we were able to make better progress in creating, discussing, and assigning codes.

Research Method: Grounded Theory
Because little research to date has been done on how users conceptualize and articulate their data needs, we chose a grounded theory approach, which is an exploratory, iterative methodology. This inductive approach seemed well suited for our purposes, as we did not start out with any particular hypothesis or hypotheses, but we knew that we had a rich data set. In grounded theory, researchers constantly ‘move back and forth’ between data collection and analysis (Bryant and Charmaz, 2007), resulting coincidentally in data refinement and conceptual categorization that leads to increasingly theoretical insight (Payne and Payne, 2004; Bryant and Charmaz, 2007).

On our first pass at analyzing the chat transcripts, we used the process of open coding and memoing (Grounded Theory Institute, 2014) to look for common patterns and to recognize and establish emerging themes. From there, we developed nascent codes and descriptors to start categorizing the data; codes were applied to relevant portions or passages of transcripts, while descriptors were applied to entire transcripts. We used the process of constant comparison (Grounded Theory Institute, 2014) to scrutinize and further develop codes and descriptors as we applied them.

During this initial phase, we communicated on a regular basis through memos and real-time meetings to discuss observations, to deliberate over the shape of the emerging coding/descriptor schema, and to consider strategies that would better focus the data set. This iterative process of collaborative inquiry—i.e., observation, analysis, deliberation, and refinement—likewise marked each subsequent phase of our investigation, as the data collection and analysis processes described below demonstrate.

While we remain in the exploratory stage of our investigation, using a grounded theory approach will allow us over time to move from coding, categorizing, and comparing concepts to building an overarching theory that we can then marry with existing literature on the topic (Grounded Theory Institute, 2014).

Data Collection and Analysis
Due to the iterative nature of grounded theory, most of our data collection and analysis processes were inextricably entwined. Initially, we collected three years’ worth of chat reference transcripts, as text files, from LibraryH3lp, our chat service provider. We then used two main tools to compile our data: FileLocator Pro, to retrieve transcripts containing data-related keywords, and TextCrawler, to remove system-generated librarian identifiers.

To analyze these transcripts, we used Dedoose, a web-based application developed to perform mixed-methods analyses in the social sciences. Dedoose allowed us to categorize each transcript using controlled descriptors—for example, to indicate whether a transcript should be included or excluded from a sample, and also to apply qualitative codes to excerpts of text within the transcripts. These descriptors and codes could then be cross-tabulated, analyzed, and visualized in various ways. In combination, these tools—FileLocator Pro, TextCrawler, and Dedoose—were extremely effective for selecting a sample of transcripts, for protecting the privacy of individuals involved, and for classifying and analyzing the transcripts within a sample.

The process of gathering data-related reference transactions, however, was a non-trivial task. Even generating a starting search strategy required careful consideration of disambiguation. For example, we quickly realized that a search for the phrase, ‘number of’, would also retrieve results where a librarian or patron mentions the call number of a book. After a few minor tweaks to minimize these mismatches, our search strategy settled on this:

- data OR statistics OR stats OR GDP OR demographics OR census OR mortality OR GIS OR quantitative OR numeric OR SPSS OR Atlas OR Atlas OR NVivo OR qualitative OR vivo OR “Data Services” OR data.services@nyu.edu OR “Data Service Studio” OR “data.service@nyu.edu” OR STATA

This limited the number of transcripts substantially, but still retrieved an immense number of transcripts that were not data-related. For example, patrons and librarians often use the word ‘data’ casually when discussing databases or information in general. Additionally, there were quite a few hits where the patron was asking for help locating or accessing a book or article that had one or more of our search terms in its title, yet the resource itself was not data-related (e.g., a quantitative study related to nursing). There were also cases where the physical space of our Data Services department was referenced, but not in regard to data needs (e.g., complaints of an unruly patron or broken computer in that area).

In order to ensure that the sample contained as many data-related results as possible, we read through the transcripts, looking for actual relevance to data, and assigned an inclusion or exclusion descriptor to each one. Even so, we ended up with 950 data-related transcripts from just one year’s worth of transcripts. So we further refined our inclusion/exclusion criteria to omit those data-related transcripts involving ‘known item’ questions, such as a patron asking for help locating a specific financial report that contained data they had found via Google. While sometimes these patrons seemed clearly interested in the data that the report contained, it was often difficult to say whether this was definitively the case, or whether they were more interested in the report as a whole. We applied these new descriptors to the sample. At this point, 633 transcripts remained, a large proportion of which still involved questions about specific databases for business and financial information.

At a loss for ideas of other wide-sweeping exclusions we could make, we made a first pass at creating descriptors and codes for the transcripts in this sample. We read through them separately, coming up with lists of descriptors/codes that seemed potentially relevant, such as which specific resources were mentioned, the general subject area of the query, how accurate the librarian’s answer was (on a numeric scale), and how satisfied the patron seemed (on a numeric scale). We then discussed our experiences as a group and quickly realized the overwhelming effort that would be necessary to apply multiple, quantitative descriptors to a sample of this size. We decided to drop nearly all of the descriptors, and instead, apply codes within the text of each transcript, indicating the presence of different characteristics, like ‘inaccurate answer’ or ‘patron satisfaction.’ This was a speedier process, and we were able to make better progress in creating, discussing, and assigning codes.
Although we were now making more progress, we discovered that the sample did not include as many juicy, in-depth data reference questions that we had hoped to explore. After a few more code-refining group discussions, we introduced a new code that indicated simply which transcripts were compelling. We focused on these transcripts, looking for patterns that might help us come up with a new iteration of our search strategy. In doing this, we were surprised by how many reference questions we received that were explicitly related to United States and international census data, and, conveniently, it seemed like these questions tended to be the more in-depth exchanges that we were after.

We completely revised our search strategy, so that it included the terms that were frequently used in these interactions:

census OR Factfinder OR "Social Explorer" OR "American Community Survey" OR "Fact Finder"4

This strategy retrieved 147 results across all three years of transcripts, although, of course, there are some caveats to the ’meaningfulness’ of this search. For example, it only captures use of the word ‘census’, so sometimes questions are included which merely involve the concept of a census or patrons may ask for known items, other than censuses, that happen to have the word ‘census’ in the title. It also relies on user and librarian understanding of when to consult a census: sometimes the user is wrong, sometimes the librarian is wrong, and our sample includes both of these cases. Furthermore, this strategy omits census-related questions where the patron’s information need was not sufficiently explored or understood, such that a census would have been an appropriate suggestion on the part of the librarian, but the transaction never got that far.

For each transcript in this new sample, we started by examining only the patron’s opening question, unnegotiated in any way by the librarian. We made observations about more easily categorizable and quantifiable aspects, like what time period was requested, as well as more qualitative, nuanced observations on the phrasing used by the patron. As before, we separately compiled lists of our observations; these ended up being extremely similar. Where there was no difference in what was observed, we created a corresponding code. Where disparity occurred, we discussed potential options and implications until consensus was achieved. We then applied this coding scheme to the transcripts.

We were interested in exploring further the qualitative aspect of the users’ questions, potentially using this to develop theories about how the users conceptualized data. In consulting the library and information science literature for other studies on how users formulate information requests, we came across an article that examined reference questions submitted to archives staff via email (Duff and Johnson, 2001). We expanded the scope of our coding beyond the patron’s initial statement of need, categorizing the overall kinds of information given and wanted by the patron, as Duff and Johnson had done.

**Preliminary Findings**

Below is a quantitative and qualitative snapshot of some of the observations and themes we have been able to extract from the data thus far using the iterative processes of coding and categorization.

**General Observations**

Not all patrons asked for ‘data’ in the data reference questions we identified. In fact, users invoked various terms to describe their data needs. Figure 1 breaks down the frequency of language that patrons used to communicate their need for data.

[Roughly one quarter of users did ask explicitly for ‘data’. Another quarter of users used alternative language that implied that they were looking for quantitative or numeric information, while a third quarter asked for either ‘information’, ‘statistics’, or ‘stats’. The remaining quarter of users asked for specific publications types that possibly contained data, e.g., journal articles, research reports, or books.]

Some patrons were very specific about temporal and geographic aspects of their data needs, while others were not. In some cases, this information was freely given in their opening statements, in others, such details emerged through a reference interview. Overall, 49% of users voiced data needs that included a specific time period; of those, 38% sought historical data or data from a range of years, while 9% sought the ‘most recent’ data available.

![Figure 1 Words initially used by patrons to describe their data needs.](image-url)
In contrast, only 4% of users indicated a specific time scale (e.g., annual, decadal). 82% of users asked for data from a specific geographic location; of those, 68% sought United States data and 27% sought New York City data. 79% of users described data needs that included a particular geographic scale; of those, 32% sought city-level data, 17% sought country-level data, and 12% sought neighborhood-level data. In many cases, it was difficult to know exactly which geographic scale a patron actually needed unless it was expressed at the most granular level. For example, a user asking for New York City data may have actually needed data on Harlem (a neighborhood within New York City), which they may have thought—correctly or incorrectly—would be findable in the city-level data set.

The nature of patrons’ data needs also varied across subject area, as Figure 2 demonstrates.

Nearly one third of all the data queries we identified were in reference to demographic data, while roughly one fifth were in relation to business, industry, and marketing data. Together, demographic and business data reference questions constituted the bulk of our data set.

Lastly, 36% of the transcripts we identified showed ‘referral activity’. This means that they had been transferred between different librarians within NYU’s LibraryH3lp system, that the librarian had consulted with another librarian during the course of the chat, or that the librarian had given the user another librarian’s contact information for follow-up. This suggests the collaborative nature of data reference as well as demand for specialized data and/or subject expertise in our sample.

Emerging Themes
Data analysis is still ongoing, but a number of themes have emerged that are worth further exploration. Although there are many interesting themes related to patrons’ question topics, librarian responses, and general characteristics of the interactions, the ones described below focus on patron behavior, and specifically on how patrons pose their initial questions to the librarian.

The Easiest/Fastest Way
The first theme describes when a patron specifies that they are not only looking for data or statistics, but specifically for a faster or more efficient way than they can devise on their own. Several examples appear below:

Patron: I’m wondering what is the most efficient way to find NY Census data from 1840-1940...I just need general numbers/demographics

Patron: hi - i’m trying to figure out how many Italians immigrated to the US at the end of the 19th-early 20th century
Patron: is there an easy way to find this?

Patron: Hello, I am trying to locate health statistics for the borough of Brooklyn from the census. Can you suggest a link? The census is a bit convoluted and I am a bit rushed.

By asking the question in this manner, the patron could be implying that they believe they have the ability to find what they are looking for if only they had enough time to do it. Along the same lines, they could also be phrasing their question this way to ‘save face’–that is, to make it seem to the librarian like they are more confident about their searching abilities than they really are. The patron could also be admitting that they know that what they need is likely to exist, but know that they lack the skills to find it.
Ask (For) A Librarian

Instead of asking for help finding data, several patrons instead asked directly for a person who might know the answer to their question. For example,

Patron: Hello–is there someone who is great with using the Census website?

—

Patron: Hi - where would I find someone who knows about Gov Docs?

—

The patrons who asked their questions this way showed a fairly sophisticated understanding of the library’s reference service; that is, they understood the concept of specialist librarians, that many data and statistics questions go beyond the realm of general reference, and that there are librarians on staff who specialize in data and statistics areas. Of course, it is difficult to know the patron’s true mindset in phrasing a question like this, but it could be read as either benevolent (indicating to the general reference librarian that it is ok if they do not know how to help with a very specialized question) or impatient (immediately asking for a specialist knowing that communicating with the generalist may not be a good use of time).

‘Am I In The Right Place?’

On the other hand, many patrons began their conversation with the librarian by admitting their inexperience with the reference service model in asking a first question about whether or not the librarian might be able to help them, or verifying what they might expect to receive from the librarian. Here are a few examples of this:

Patron: I’m looking for information regarding United States annual steel production as far back as possible, to present
Librarian: ok
Patron: would you be able to help me find that info? perhaps recommend some material

—

Patron: Hey I have to find some figures on topics based on cities, if I were to tell you some of these topics do you think you can give me a hunch on where to start or which databases would be helpful?

—

Patron: I have a question about citing US census data?
Patron: (I’m not sure if that’s something you could help me with)

Interestingly, this patron could potentially have the same spectrum of intentions as the savvier patron who asked for a librarian above. By expressing doubt about whether the librarian can help, they again make it ok for the librarian to say they do not know how to help (or to give basic help or make a referral) and it potentially saves time by making sure they are asking the question to the right person in the right place.

Authority

Another common theme arose, relating to the authority and reliability of sources the patron had already found—a theme that will be unsurprising to anyone who does any reference or information literacy instruction. For example:

Patron: Hi, I need an academic source that establishes the years for all living generations. Could you help me find a reputable source?

—

Patron: hello! im looking for demographics on southern brooklyn (birth rates, sex, age population). we are not allowed to use wikipedia as a resource

—

Patron: i can’t seem to find what i want
Patron: is indexmundi.com a reliable source?

—

Patron: I am researching the recents stats of homelessness in NYC
Patron: how can I find accurate numbers?

This theme suggests a more substantial knowledge gap for the patron—the lack of ability to evaluate the reliability and authority of a source—but also the wherewithal to acknowledge this gap and ask for help. It is difficult to tell from most chat transcripts whether these patrons were interested in authority for the sake of an assignment (i.e., their instructor told them they can only use authoritative sources) or for the sake of having reliable data for their own projects or needs, but it is likely that both types are represented.

‘Where’ vs. ‘How’

Another interesting distinction that emerged was that some patrons ask for ‘where’ to find the data they need while others ask ‘how’ to find it. For example:

Patron: I was trying to find demographic information from 1980 to 1990 for Far rockaway, ny
Patron: where should I look

—

Patron: Hi, do you know where can I find the total number of college students in specific cities

Versus:

Patron: Hello, I need to find cities in US where people need to use public transportation a lot
Patron: Do you know how can I find the data?

—

Patron: i want to find the revenue number of taobao.com, an ecommerce website in China
Patron: could you show me how to find the numbers? thank you

While this could simply be a result of different manners of speaking (rather than something deliberate and worth analyzing), it could also reveal clues into the way different patrons are thinking about their data needs and questions. Both patrons seem to assume that the data they need exists, but the one who asks ‘where’ also seems to believe that once they know where to look, then the process of extracting or accessing it and understanding what it means will be easy or at least doable. This patron could be a more experienced data user, or could be overestimating their abilities. The patron who asks ‘how’ is acknowledging that they do not know how to approach searching and possibly also does not know what to do.
once the desired data are found. Looking at whether a patron asks ‘where’ or ‘how’ may also tell us something about where the patron is in the research process, for example, if they are looking for data or statistics to support an argument that they have already made, or if they are in a more exploratory stage.

**Unanswerable**
Finally, we will explore a broader, more complex category of patron questions that we have chosen to classify as unanswerable for some reason or another. This does not mean that the question is not legitimate or should not have been asked, only that the way that it was asked makes it impossible to answer at face value. Essentially, these questions are ones that require a good reference interview on the part of the librarian, and looking closely at the original phrasing of the question gives us interesting insight into how the patron was thinking about the information need and approaching it for the sake of the librarian. There are several flavors of the unanswerable theme, which are discussed after each example, below.

Patron: Hello - I have been searching a statistic for two days and I have been unsuccessful and running out of time. Can you help me?
Librarian: Hi there
Librarian: I am trying to find the uninsured rate (for healthcare) in Canada - and cannot for the life of me find it
Librarian: I know Canada has universal health care but I need a solid statistic within the past 5 years of those citizens that are uninsured.

In this case, the patron is asking for something that they admit should not logically exist: if Canada has universal health care, then there should not be any uninsured Canadian citizens. Even so, the patron clearly has an information need; it is reasonable to assume that they are aware of this logical fallacy, so the librarian’s job is to help clarify that need and then help fulfill it. This is, in fact, what happened over the course of this chat conversation. It could be that the patron had spent sufficient time on this project such that when they asked the question, they forgot that the librarian would not have the same context to understand what was meant by this query.

The example above also hints at the patron’s challenge in operationalizing concepts into variables that are likely to exist and be available. This was observed many other times too, for example:

Patron: Hello I’m currently working on a project about the changing face of Jersey Street in New Brighton, Staten Island. How would you advise that I find out the culture of crime in that area from 1950s through now?
Librarian: Hi there
Librarian: Are you looking for books? articles? statistics?
Patron: stats please

In order to find statistics on the ‘culture of crime’ in a certain area, this patron will need to decide how the concept should be defined and measured first.

Patron: how would I find the specific ethnic breakdown and class breakdown of East Los Angeles? I need information on that specific region
Librarian: In East Los Angeles.

Likewise, while there exist some standardized definitions for collecting data on ethnicity (though these can and should be scrutinized), there are no similar standards for data on ‘class’ in the United States. This patron will need to clarify what they mean by ‘class breakdown’ before they can find statistics about it in East Los Angeles.

Other users asked for things that were simply unlikely to exist or be available publicly or through library databases, for example:

Patron: I need statistics for US tomato consumption in 1840s, 1850s...thru 1900. USDA stats start in 1886

Many reference librarians will read this patron’s statement as a successful search: the patron identified the correct authority most likely to have the data if they exist; however, since that authority does not have them, the answer is that those data almost certainly do not exist. There could be some additional discussion about proxy variables or other creative places to look and maybe this patron’s need could still be satisfied, but the interesting part is the difference in how the patron thought about this question versus the way a professional librarian would.

Most of the queries that fall into this category also raise questions about what background work the patron has already done and whether they might be better served by looking for books or articles instead of data sources.

**Limitations and Next Steps**
We acknowledge, of course, that our approach has limitations. While chat transcripts allow us to look back at reference transactions in a way we never could with in-person reference, we also do not have any feedback about the experience from the patron or the librarian. As a result, it is very difficult to truly know what the patron really wanted, or whether the patron or librarian considered the interaction successful. Furthermore, the concept of a successful interaction is complicated by the fact that user satisfaction or dissatisfaction does not necessarily equate to a correct or complete answer from the librarian. For example, is an interaction successful if the librarian determines that the desired data exist, but only in PDF format, and then the user leaves discouraged? Or if the librarian gives an answer that is wrong or incomplete but the patron is happy with the answer? Additionally, the set of transcripts we extracted may be incomplete, because it is difficult to identify transactions where neither the patron nor the librarian recognized a data need, which may be among the most interesting interactions.

There are many additional themes in the chat transcripts in our data set; this investigation is a preliminary exploration of how patrons ask data-related questions. More themes—and their relationships to one another—will be discussed in future publications.

A grounded theory approach suggests that the next phase of this project will be to begin exploring the relationships between themes and determining what this data set is a study of (Grounded Theory Institute, 2014). From the themes already uncovered, we have several pressing questions:

- Are these themes specific to census-related questions? Are they even specific to data-related questions? Or are they more generalizable to all chat reference?
- Is there a relationship between any of these themes and the overall success or failure of the reference interaction?
Once we have built a theory or theories from the data, the final step will be to integrate them into the established literature and articulate how our work moves the conversation forward, possibly adding to a growing body of knowledge about the librarian’s role in supporting the data lifecycle.

In addition to the theoretical advantages of understanding our users, there are practical aspects of this inquiry as well. This project gives us a rare opportunity to look closely at some of the problems our users and librarians are having with data in reference transactions and to think about how we can improve our services for the benefit of all. In better understanding the kinds of queries we receive, and the ways data needs are conceptualized and articulated, we hope to build better data research guides for our patrons and improve the training, scripts, and guides available to the librarians staffing the service.

One clear way to improve service is to offer training to library staff on how to use open-ended questions during the data reference interview. As evidenced by questions classified within the ‘unanswerable’ theme, users often have an incomplete understanding of how to operationalize concepts into variables that could be found in existing data sources. Training that allows staff to practice asking the kinds of open-ended questions that will help users and librarians move toward a shared understanding of what the user needs, and what exists, will translate into more effective data reference interactions.

Our analysis also shows that users struggle with questions related to the reliability and authority of data sources. This could be communicated efficiently through an online guide showing the who, how, and why of data creation, collection, and distribution, as well as strategies for evaluating sources. Making this kind of information available allows librarians to more easily seize a teaching moment, and enrich and expand the learning experience beyond the immediate data reference interaction. These guides are especially valuable because they make it easier for generalists staffing the service to convey specialized information. These are just two possible ways to improve service based on our preliminary findings. As demand for secondary data grows across academic disciplines, strengthening the data reference piece of a larger reference program that is staffed by specialists and generalists alike ensures the future health and relevance of academic reference services.

References


Notes
1. Margaret Smith is Physical Sciences Librarian at New York University and can be reached by email at margaret.smith@nyu.edu.
2. Jill Conte is Librarian for Sociology, Psychology, and Gender & Sexuality Studies at New York University and can be reached by email at jill.conte@nyu.edu.
3. Samantha Guss is Social Sciences Librarian (for Political Science, International Studies, Geography & the Environment, Government Information, and Data & Statistics) at the University of Richmond and can be reached by email at sguss@richmond.edu. She was previously Data Services Librarian at New York University from 2009-2014.
4. American Factfinder is the United States Census Bureau’s online tool for accessing data. The American Community Survey is a demographic survey that complements the United States Decennial Census. Social Explorer is a commonly used commercial database that repackages U.S. Census and other data.
Digitization, Data Curation, and Human Rights Documents:

Case Study of a Library-Researcher-Practitioner Collaboration

by Amy Barton, Paul J. Bracke, Ann Marie Clark¹

Abstract
At Purdue University Libraries, a project involving the digitization of Amnesty International Urgent Action Bulletins from 1974-2007 combined the strengths of political science and library science researchers. The political science research was centered on transnational human rights advocacy and legal instrumentation changes over time, while the Libraries’ research related to data management, data lifecycle and curation, metadata, and collaborative research modeling. The conceptual framework of this case study is rooted in the literature on embedded librarianship and lifecycle models of data curation. We investigate the intersections and alignments between scholarly workflow and curatorial workflow, and the implications of these intersections and alignments in collaborative research and curatorial lifecycles. The case study also examines how library resources supported research, and how library science and political science experts collaborated in research through the development of a conceptual model. A research collaboration model was developed specifically for the human rights texts project, but was then generalized to be applicable for a variety of practitioner-librarian collaboration projects. The research resulted in data production, data curation, data management, data publication, and scholarly communication and dissemination.

Keywords: data curation lifecycle, metadata, human rights, Amnesty International, research collaboration, digitization, archives

Introduction
Partnerships among librarians and faculty members that develop ways to preserve and create digital access to research information have the potential to open up new avenues of teaching and inquiry for both faculty and librarians. A project at Purdue University to create a digital research collection of human rights documents has piloted this sort of innovation and collaboration. By its nature, the project required close communication at key points to make the most of faculty and library expertise. This paper explores the process of creating a digital collection of international human rights documents in a way that is integrated into the research workflows of faculty.

Academic Libraries have been exploring new roles in recent years that improve engagement and more tightly couple the activities of the Libraries with the activities of their users and institutions. This has been manifest in several ways. Significant attention has been paid to better integrating digital collections with discovery and linking technologies, and ‘getting into the flow’ of faculty and students (Dempsey 2012). There has also been considerable focus on revitalizing liaison librarian roles as a way of developing stronger relationships and partnerships with faculty, particularly in support of information literacy, scholarly communication, and support for digital scholarship (Auckland 2012; Jaguszewski & Williams 2013; Kenney 2014). Information literacy and informed learning, for example, imply a deeper level of integration into the curriculum than bibliographic instruction efforts of the past (Jaguszewski & Williams 2013).

Libraries have also been active in recent years in exploring opportunities for supporting scholars in their research processes. There has been considerable interest among research libraries, for example, in supporting digital humanities and e-science services. These often involve partnerships between libraries and researchers to find ways to better integrate library collections, library science expertise, and library services into research processes, and to better integrate primary research outputs, such as datasets, into curatorial processes (Brandt 2007). The digitization project for the Amnesty International Urgent Action
Bulletins collection at Purdue University Libraries, referred to below as the Human Rights Texts project, developed out of such a partnership between faculty and the library. The goal was not merely to develop information products that are of general use to scholars, but also to align the development process with lifecycle models of data use and scholarship support.

Conceptual Framework and Research Questions

The conceptual framework of this case study is rooted in the literature on embedded librarianship and lifecycle models of data curation. Embedded librarianship is an emerging model of librarianship in which librarians apply expertise in information organization, management, and use in the context of library users — embedded within teaching, research, or clinical environments (Schumaker & Talley 2009, p. 8). Rather than a passive model in which librarians react to the needs of their constituents, embedded librarianship presents an active model in which librarians are partners in addressing instructional or research needs (Kesselman & Watstein 2008; Carlson & Kneale 2011; Clyde & Lee 2011; Rudasill 2010). While many librarians, especially those at Purdue University Libraries, have focused their efforts in support of information literacy and instruction, there is increasing interest in developing partnership models for research collaborations (Brandt 2007). Numerous discussions in the literature engage in how librarians can develop relationships with scientists and social scientists related to data management, and describe pilot programs to do so (Brandt 2007; Garritano & Carlson 2009; Walters 2009). There is also a history of librarian partnerships in the digital humanities, often through the creation of digital humanities centers, in which scholars may receive support for research projects. For example, assistance with digitization or text analysis (Vandegrift & Varner 2013; Posner 2013; Gold 2012; Svensson 2010; Zorich 2009; Zorich 2008).

This study also is framed by lifecycle models of data management and research collaboration. There has been significant interest among librarians and others involved in data management in recent years about data curation lifecycles, a phrase that refers to how data are acquired and cared for throughout their production, acquisition, use, and preservation (Carlson 2014). One of the most cited of these models is the DCC Curation Lifecycle, developed at the University of Edinburgh’s Digital Curation Center (DCC). This model is designed to aid the planning of curation and preservation activities, and the long-term continuity of access to digital materials (Higgins 2008). Accordingly, it is largely focused on activities directly related to curation and preservation, and presents a model to which more granularly-defined local practices could be mapped. The model, seen in Figure 1, places the digital object at the center, with central rings representing curation actions applicable across the entire lifecycle of the object and outer rings representing sequential or occasional actions that are performed upon the data. While this is a robust, generalized framework for planning, it is also a framework that was created with a specific purpose — digital curation. While curatorial activities and responsibilities were well represented in the model, the scholarly processes that drive the creation and use of digital objects are under-represented.

Other models (DDIAlliance.org 2013; Green & Humphrey 2013; Vardigan et al. 2008) have inverted this focus on developing a
model of data-based research that feeds curatorial processes.

Taking the Data Documentation Initiative (DDI) Combined Lifecycle Model (Figure 2) as an example, the lifecycle begins with the genesis of a research study and progresses through a number of stages of data preparation and use, with a step for data archiving and distribution. This view is researcher-centric and underrepresents curatorial processes.

Designing a practice to integrate two different lifecycle models presents a challenge. For the project presented in this case study, the practice must serve both research and curatorial goals: first, by providing ongoing research access to primary source documents, facilitating qualitative coding of the documents and ultimately the creation of qualitative and quantitative data sets; and second, by simultaneously incorporating the curatorial steps required to create the lasting collection of primary source documents, a research database, and data that are appropriate for long-term curation and preservation. The researcher and the library units involved in this project envisioned mutual benefit in building a process that could use research steps to facilitate curation and, in addition, employ curatorial techniques to facilitate research. Accordingly, this case study sheds light on a question central to research on integrated models of the data lifecycle: can digitization processes be designed in a manner that feeds directly into analytical workflows of social science researchers, while still meeting the needs of the archive or library concerned with long-term stewardship of the digitized content?

Answering this question, from the standpoint of an academic library, leads to two subquestions. First, what are the intersections and alignments between scholarly workflow and curatorial workflow? Second, what are the implications of these intersections and alignments on research and curatorial lifecycles?

The Project

Since the mid-1970s, as an advocacy technique in support of human rights, Amnesty International (AI) has produced periodic, almost daily Urgent Action (UA) Bulletins (Rydkvist 2013). These Bulletins are shared with members of its UA network, who are asked to write direct appeals on behalf of individuals whom AI believes to be at risk of human rights violations. Examples have included action on behalf of people who may be tortured, disappeared, or detained illegally, or severely mistreated while in detention. The Bulletins themselves are one- to three-page informational alerts advising members to write quick messages directly to officials in violating governments. In addition, the Bulletins advise members on how to raise issues of concern in each specific case. Concerns may range from protecting the affected person’s physical well-being to other forms of compliance with relevant principles of human rights protection.

The UA Bulletins collection provides a detailed record of a major human rights organization’s transnational advocacy on behalf of individuals over more than four decades. Data can be drawn from the documents to serve as indicators of how human rights concerns changed over time, as well as the nature of human rights threats in different countries and different periods. The documents also provide a unique window into the mobilization of the human rights movement during a crucial time period in the development of international human rights law (Clark 2001). The digitized collection will be of use to researchers, as well as journalists and attorneys who can refer to the documents as primary evidence from the historical record.

The project to digitize the AI UA Bulletins was initiated for several reasons. The researcher, a faculty member in Purdue’s Department of Political Science (Principal Investigator [PI]), had served for a number of years as a volunteer on Amnesty International-USAs (AI-USA) Archives Advisory Committee. In this capacity, she learned of the need to preserve the early records of AI-USA, the United States branch of the global human rights organization. The AI-USA offices, however, housed additional documents with significant research value that were publications of AI’s International Secretariat (AI-IS), the organization’s international headquarters in London. These documents included the UA Bulletins and associated documents as they had been processed for distribution to members of the network in the United States. While the preservation needs for documents created by AI-USA were being addressed with the establishment of its organizational archives at Columbia University in 2007, digital preservation of the UA documents originally created by AI-IS were ultimately deemed out of scope for Columbia’s collecting practices.

Between 2007 and 2011, the researcher engaged in extended conversations with staff from AI-IS, AI-USA, and the Amnesty International USA Archives (Columbia University, n.d.) about developing joint digitization and research as a stand-alone project related to the UA Bulletins processed and issued by AI-USA. After permission to pursue funding for digitization and research with the documents was approved by AI-USA, the researcher contacted the Purdue Libraries. With the Libraries, discussion centered around how to prepare the UA Bulletins as a publicly available collection in conjunction with the PI’s research process, as well as how best to incorporate them into the Libraries’ digital collections. A funding proposal was developed that sought to combine the strengths of the librarians’ archiving, technical, metadata and research data expertise with the researcher’s scholarly human rights expertise, and contacts at AI-USA and AI-IS, who saw the benefit of a digitized collection of these particular documents. AI-USA agreed on the general outlines of the project as envisioned and agreed to share the documents. A competitive grant from the Purdue University Office of the Vice President for Research funded the project (Clark & Bracke 2012). The proposal was to combine library standards for digitization and digital collections, as well as additional researcher- and practitioner-driven metadata and coding strategies. The collaborative research would result in a searchable, e-Archive primary source collection that would also benefit the human rights organization, as well as a human rights dataset for researcher use, with the potential for expansion into a numeric dataset compatible with other international data sources.

Working with AI-USA required extensive consultation, both before and after funding was secured, to ensure that each of the parties’ needs and interests, and sensitive information concerns, would be taken into account. Variation in the timelines and organizational approaches of each of the three parties (library, researcher, and human rights organization) posed further challenges. While these logistical issues slowed progress as the project got off the ground, in retrospect it was a necessary feature of the collaborative process.

Research Process

Processing Workflow

Archival best practices were established for the handling of the documents during digitization and post-processing reassembly. This ensured the documents could be sent along to their...
permanent home at the Amnesty International USA Archives at Columbia University in a good archival state.

The digitization process (Figure 3: 1.a Digitization) was standard, with the exception of the creation of the Optical Character Recognized (OCR), full text PDF derivative. High resolution and high quality OCR were necessary for accurate topic modeling by the PI and her research team. NVivo software (QSR International 2014) was used by the PI to code the digitized documents. The quality of the processed PDF documents was crucial for text recognition and topic modeling in the NVivo software for accurate and comprehensive coding. Digitization was driven by the need for processed documents to feed into the coding progression. In this scenario, it became a challenge to keep up with the need for high quality derivatives, which were time and labor intensive to produce, and the need for documents to code. Ultimately, these activities aligned such that the digitization and coding co-occurred in real time (Figure 3: 1.b Coding).

**Data and Metadata**

AI-IS and AI-USA each provided a data source documenting records of UA Bulletins and associated documents from 1974-2013. These sources had been independently maintained, resulting in differences in the underlying data. Each contained some unique information, and some records were incomplete in one source or the other. Combined into a single metadata master file along with technical metadata captured during digitization (Figure 3: 1.c Data Merge), the two sources provided comprehensive information such as the organization’s internal document numbers; dates and countries addressed; document types; and some details of the case. The combined file is being used as master metadata file for the public-facing primary source digital collection and as a basis for a research database for the PI (Figure 3: 2. Metadata & 3. Collection).

In adherence to library science best practices, we identified three human rights authoritative vocabularies and used them, in combination, to create a controlled vocabulary for the project that would supplement ad-hoc keywords assigned by AI-IS and AI-USA in each data source. The combined authoritative vocabularies include HURIDOCS’ ‘Micro-Thesauri: A Tool for Documenting Human Rights Violations,’ Witness Media Archive Topic Terms’ published by WITNESS, a human rights NGO, and the ‘Universal Human Rights Index Research Guide’ produced by the United Nations Office of the High Commissioner for Human Rights (HURIDOCS 2010; United Nations Office of the High Commissioner for Human Rights n.d.; Witness n.d.). The PI and her research team combed the vocabularies to identify semantics that closely align with her research interests as well as the descriptive terms used in the two Amnesty International data sources. The development of the controlled vocabulary was used to develop coding structures for content analysis within NVivo. This allowed the research team to create an export file within NVivo (‘node extracts’) that listed each code applied to a document. These node extracts were,
in turn, added to the metadata master file by the Libraries. This aligned indexing terms between the researcher’s NVivo research files, the master metadata file, the digital collection, and the research database.

The controlled vocabulary we developed will be shared with AI for possible applications in their own data sources. Interoperability and metadata exchange opportunities in any future projects with AI are more feasible if the concepts and terms are standardized in all parties’ data sources.

**Findings and Discussion**

This research project was intended to explore whether digitization processes can be designed in a manner that feeds directly into content analysis workflows of social science researchers, while still meeting the needs of the archive or library concerned with long-term stewardship of the digitized content and data. The authors sought to understand the intersections between scholarly workflow and curatorial workflow and their implications for research and curatorial lifecycle. The project illuminated both opportunities and challenges for advancing such collaborations.

**Intersection of workflows**

One of the key findings of the project was that, while some of our ideas about integration were correct, there were significant difficulties in aligning workflows between library and researcher from a scheduling point of view. Areas of success in aligning workflows included developing shared standards for file naming, application of the controlled vocabulary, and extraction of code structures from NVivo to provide subject enrichment of descriptive metadata being processed from AI-IS and AI-USA data sources. The successful extraction of code structures for descriptive purposes demonstrated that qualitative coding processes can be leveraged in metadata creation. Despite these points of overlap between the two, a fundamental challenge was achieving alignment in the context of a grant-funded project with time constraints.

The timelines associated with a grant-funded project required both parties, library and researcher, to begin their work as soon as possible. The first issue that arose was the need for a formal agreement between Purdue and Amnesty International prior to shipping documents for scanning. While this process went relatively smoothly, it did result in a delay of several months in having the ability to begin scanning of documents. While graduate assistants working for the researcher were able to begin their coding process with newer documents, available in born-digital formats from AI-IS (Figure 3: 1.b Coding) that could later be matched with scans of the same documents from AI-USA, the delay in beginning the scanning process would prove problematic later in the process as it was difficult to maintain a steady supply of scanned documents for coding. The second issue that arose, from a project planning point of view, was that initial planning was done based on photographs of sample documents from the collection taken by the PI during a visit to the AI-USA archive. Once documents were received, it became apparent that OCR processes would be slower than anticipated due to document condition. A third challenge resulting from the grant-driven timeline was with the use of student labor in scanning. When the documents were...
received for scanning it was mid-semester, making the hiring of new student workers more challenging than it would have been at the beginning of the semester. Additionally, there are inconsistencies in productivity inherent in the use of student labor. Students are often less available for work, at short notice, during exam periods and during semester breaks.

**Intersection of research and curatorial lifecycle**

Working through the phases of the research project, we discovered that the intersections and alignments of research and curatorial lifecycles include collaborative research, the alignment of research and curatorial processes, and the development of a new research partnership model. Due to the way the project unfolded, with delays, digitization challenges, and labor and time constraints, much of the research had to be done in real-time. The research model worked well for the project. The PI and her research team, and the faculty from Purdue Libraries and their digitization team, were able to identify research problems and processing issues that impacted research, then work together to produce the intended research outcomes.

The Libraries supported the project with services such as digitization, digital content management and digital collection development. However, our experience in this project went beyond typical services. Two Libraries faculty collaborated with the political scientist in research that resulted in the development of a dataset (populated master metadata file) that further facilitated the development of a research database and a public-facing, primary source digital collection (in development). We felt this was a fairly unique situation in that the Libraries faculty was engaged in research alongside the political scientist in the context of research data development, metadata application, data management and curation. Below, we present a model that explores the process and components involved in a successful research collaboration based on the Human Rights Texts project.

The conceptual model (see Figure 4, page 31) represents scholarly collaboration between political science and library science experts, along with Libraries services and resources, to conduct research that resulted in metadata and data development, data curation throughout the data lifecycle, digital collection development, and scholarly outputs and dissemination.

In the model, the dark blue represents the research continuum from start to completion. The light blue circles on the left of the model represent Libraries services and resources that supported the research process. The circle in the middle represents political science and library science expertise converging in real-time research. Finally, the white circles on the right of the model represent outcomes.

The seeds of the project were sown by the PI as early as 2007, but ‘Project Development,’ in the first dark blue circle, began after the Libraries and the PI had discussions and committed to the project. Funding was sought and awarded. The next point on the continuum, ‘Domain Expertise,’ the second dark blue circle on the continuum, overlaps the first set of light blue circles representing...
Libraries services and resources. At this point, Archives and Digital Programs established standards and workflows for digitization, while the Libraries established data management and curation standards and workflows in consultation with political science and library science experts. This is also the place where the PI and metadata specialist developed the controlled vocabulary to be used throughout the project. In the model, the PI and Libraries faculty working with the Libraries IT to develop the research database based on the data-populated master metadata file and the data sources merge workflow. The Libraries faculty and the PI come back together in the middle circle to conduct research - interacting as researchers, addressing real-time problems, and producing research data. Here again, the continuum undulates to represent the PI evaluating her coding results on a subset of data, and the Libraries faculty working with the Libraries IT to develop the research database based on the data-populated master metadata file. Finally, the PI and Libraries researchers converge once again to evaluate the outcomes of the project: the refinement of the ---research database; the development of the digital collection; long term data preservation and curation; scholarly communication via dissemination of research results; and the publication of the raw dataset through the Purdue University Research Repository (PURR).

Having described the development of a research collaboration and partnership model for a very specific research project, we now explore how the model could be generalized and applied to other academic libraries and research collaborations, inclusive of libraries services and expertise. The generalized model (see Figure 5, page 32) represents on a research continuum from start to completion the major components of the process that can provide guidance to the process of establishing a Libraries-domain expertise research collaboration.

We previously discussed in detail how this conceptual model applied to a specific research project. Taking a step back, one can see the three major components of this model:

- research project development,
- applied research collaboration,
- tangible and intangible research results, which includes scholarly impact and dissemination.

The research project development component includes relationship building; project development and planning; identifying domain expertise combined with library expertise to address the research problem; and the library services and resources that will support the research continuum from start to completion. The applied research collaboration component occurs after the project planning, workflows, and processes have been established. In this component, combined domains apply research methods to address the research inquiry and to produce data. This component leads into the last component, research results. This includes the intangible results - solidifying relationships and continued collaborations - and the tangibles - data production, data curation, data management, data publication, and scholarly communication and dissemination.

Based on the successful outcomes of the Human Rights Texts research project, further exploration of this model and the application of this model - perhaps even a component to operationalize the model for project management - will be pursued.

References


QSR INTERNATIONAL. 2014. NVivo 10 for Windows. Published by QSR International Pty Ltd. Available at qsrinternational.com.


First Forays into Research Data Dissemination: A Tale from the Kansas City Fed

by San Cannon and Deng Pan

Abstract
The Federal Reserve System has a longer tradition of doing economic research than disseminating data from economic research. Each of the 12 Reserve Banks and the Board of Governors have research departments that together publish nearly 1,000 working papers and journal articles annually. Unfortunately, researchers have not often made the data from their papers publicly available until recently. A new program at the Federal Reserve Bank of Kansas City aims to correct this imbalance and make such data available for reuse in other research. As a pilot participant in a new dissemination platform, we have educated economists, built metadata specifications, recruited contributors, collaborated with technology and legal staff, and coordinated and built coalitions across multiple functions at our institution and others. This paper outlines the challenges faced and obstacles overcome as we created the infrastructure and workflow and took steps toward making the publication of research data a regular part of the research life cycle.

Keywords
Data dissemination, metadata, usability testing, research data management

Introduction and Background
As researchers are deluged with data, the need for them to share or disseminate their research data widely may not seem pressing—there’s plenty to go around. (The data deluge, 2010). Researchers now have more choices when compiling data to support their research, and reusing other researchers’ data is an important option. A wide range of available data—and advances in analytical capabilities and processing power—means researchers can replicate or build upon a broader variety of research than was possible in the past. Recent mistakes in (and fabrication of) research data have only highlighted the importance of research replicability and the sharing of research data.

The Federal Reserve Bank of Kansas City, like other Reserve Banks and the Board of Governors in Washington D.C., conducts research to support its monetary policy mission, contribute to the safety and soundness of banks, and promote financial stability. The Bank shares its research products with policymakers, other researchers across the System and in academia, and the public. Increasingly, this research requires analyzing vast quantities of data and employing substantial computational resources to address increasingly complex questions.

In Kansas City, more than two dozen researchers and research associates produce about 50 research products (journal articles, working papers, etc.) each year. They are part of a larger community of more than 750 researchers across the System who produce nearly 1,000 such works each year. As the need to acquire data inputs to this research has increased, so has the pressure on support staff and the budget. To help alleviate some of the pressure, the 12 Reserve Banks and the Board have been formally collaborating to acquire source data as research inputs.
While the System has established a set of services to bring in data, there are no such services for pushing out data once the research is complete. The Federal Reserve currently has no coordinated approach to preserving and disseminating research data across the System, and differences in strategies and resources have precluded a consolidated approach. Other academic domains face similar challenges (Borgman, 2012).

**Challenges and Opportunities**

Setting up a repository or an archive, or defining a workflow to support data preservation or future dissemination, are not just technology decisions. For many disciplines, these activities require a fundamental change in researchers’ perceptions of the research process. For most researchers, the ultimate goal is publication; Fed economists are no exception. All data-related work is simply in support of that goal: ‘Time and money spent on documenting data for use by others are resources not spent in data collection, analysis, equipment, publication fees, conference travel, writing papers and proposals, or other research necessities’ (Borgman, 2012). Effecting change in such circumstances is difficult but not impossible, and research funders may lead the charge. For example, some funding agencies now require publicly funded researchers to make their underlying data available to the public. Although these requirements do not affect Federal Reserve researchers, they will undoubtedly help change the culture of empirical research as a whole (Arzberger et al., 2004).

Even without funding requirements, a few banks in the Federal Reserve System have begun considering how to disseminate their research data sets. The Federal Reserve Board, for example, publishes data for select working papers along with the papers on their website. The data are being disseminated, but researchers must know with which paper they are associated and must then go to that page. The Federal Reserve Bank of New York takes a similar approach but is also compiling a separate page for such data. The Federal Reserve Bank of Kansas City, however, does not currently disseminate research data sets on its public website. For many Reserve Banks, the major means by which research data is disseminated is individual requests to the author. This process, though widespread even in the academic community, is taxing on the author and does not encourage broader reuse.

A coordinated approach to data acquisition in the Federal Reserve System began in late 2011 and was aided by the creation of a data librarian role in each of the Reserve Banks. The Federal Reserve Bank of Kansas City took research support one step further by creating the Center for the Advancement of Research and Data in Economics (CADRE) in early 2015. CADRE’s mission is to support, enhance, and advance data or computationally intensive research in economics. Bank leadership identified research data preservation and dissemination as important support functions for CADRE. As these functions were being designed, the Kansas City Fed was offered, and accepted, an opportunity to join a pilot program to help provide use cases and product suggestions for a new research data dissemination platform being developed by a nonprofit academic partnership outside the Federal Reserve System.

**Becoming a Pilot Participant**

The new publication platform is designed for researchers to submit data directly for dissemination. The workflow in the platform allows for two stages: an initial submission by the researcher and a curation step to verify, edit, update, or clarify the contents of the initial submission before publication. To participate, CADRE needed to specify the details of each stage in the workflow and build use cases for our research community. To do so we considered the following questions:

• What kind of collections? In the dissemination platform, a collection is a group of data sets for which similar policies and access controls can be set. Because metadata and access rights are controlled at this level, we created a public data collection and a restricted access data collection based on our evaluation of the data sets that had been used by researchers in the Federal Reserve System. We needed to be able to store data to which access was limited and test if the access controls worked sufficiently to meet our information security requirements. We also wanted to test the submission workflow and applicability of common metadata across the collections.

• What kind and size of data files? Although the platform was built to accommodate very large files stored in a central location, the typical data files in the pilot range from 1MB to 5GB, and are stored on the researcher’s desktop or on the Kansas City Fed’s high-performance computing cluster.

• What kind of workflows? In the initial phase of the submission workflow, submitters fill in required metadata fields to describe the data sets and then assemble data files. In the subsequent curation phase, curators review and possibly modify the metadata or files before approving or rejecting the submission.

In addition, we had some practical questions about how the pilot itself should proceed.

• Who would be the testers? We expected four to six test users, mostly economists and their research associates at the Kansas City Fed, to start this pilot and provide initial feedback. We also hoped to expand the test-user base to include users in a few other Reserve Banks as well as their co-authors at academic institutions.

• Who would be the curator? While CADRE had plans to hire a Data Curator, the position had not yet been filled. The curator role was thus temporarily filled by two staff (specifically, the authors).

**Metadata Creation**

The most critical decisions for this pilot involved choosing which metadata fields the data set submissions should capture. Opinions vary on how much information is sufficient to adequately describe any item. We investigated several existing specifications to evaluate what others view as necessary information for finding or discovering data.

First, we looked at the requirements for depositing a data set at the University of Michigan’s Inter-university Consortium of Political and Social Research (ICPSR), which maintains a data archive for social science research data. Although our data-dissemination pilot is not
meant as an archive, many of the issues for data discovery are the same. The submission process for ICPSR doesn’t explicitly require any fields, but we believe Title, Principle Investigator, and Description are the minimum metadata fields that are practicable. In addition, ICPSR requests another dozen or so categories of metadata ranging from mode of collection to geographic coverage. Because the ICPSR archive handles a large number of data sets that are primary data collections, many fields, such as response rate and weights, rarely apply to data used in Federal Reserve research projects.

Next, we reviewed one metadata schema specifically designed to help users discover data. For Federal agencies that provide data, the executive order OMB 13-13 specifies a particular metadata schema to catalog data assets. The metadata elements defined for that order are published as part of Project Open Data comprise 12 required fields, some common to other specifications (for example, title, description, keyword) and others specific to government agencies (Bureau Code, Program Code). In addition, six fields are required by the schema if applicable, including license, rights, and spatial and temporal metadata.

Finally, we examined the metadata requirements that an internal workgroup had developed to catalog data assets across the Federal Reserve System. This unpublished specification lists more than 30 metadata elements including 14 mandatory elements. Some items are common to other specifications (such as name or description), whereas others describe access restrictions (for example, security classifications).

After carefully considering the user burden for metadata entry, we decided on the following metadata elements.

<table>
<thead>
<tr>
<th>Required</th>
<th>Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set name</td>
<td>Additional information</td>
</tr>
<tr>
<td>Contact author name</td>
<td>Key words</td>
</tr>
<tr>
<td>Contact author information</td>
<td>Journal of Economic Literature Classification</td>
</tr>
<tr>
<td>Description</td>
<td>Geography</td>
</tr>
<tr>
<td>Update information</td>
<td>Unit of observation</td>
</tr>
<tr>
<td>Category</td>
<td>Date(s)</td>
</tr>
<tr>
<td>Frequency</td>
<td>Documentation</td>
</tr>
<tr>
<td>Access restriction</td>
<td></td>
</tr>
<tr>
<td>Security classification</td>
<td></td>
</tr>
</tbody>
</table>

We also added three metadata fields to the curation workflow: file type(s), file size(s), and article information. The file specification information is important both for technical staff managing the file space and for users who initiate a download. Incorporating these into the curation workflow rather than the submission process reduces burden on the depositor and allows for possible changes to file types and sizes (through compression, for example) before the data are published.

Infrastructure Choices and Challenges

Once the interface and workflow were developed to our specifications, we performed initial testing before involving our users. We faced early infrastructure challenges; specifically, getting the dissemination platform to work well with our environment. The dissemination platform is meant to interact with our existing file system and storage infrastructure. We had some difficulty getting the file system and publication platform to interact smoothly. Because we need to allow access by outside users, the file storage and platform live in an external zone of the Kansas City Fed’s intranet. This positioning made it easier for external users to get to the files, but made it more challenging for internal users to load the files from their desktop to the endpoint. The challenge was not insurmountable, but it did frustrate us as we worked through a typical user experience.

Another major infrastructure difficulty involved authentication and identity management. As part of a pilot project, we needed to have accounts on the platform infrastructure. These identities were then used to manage access to the collections through group definitions. For our initial work, adding individual users to the appropriate roles and access groups was fairly straightforward. However, in planning for a more robust long-term implementation, we do not want to maintain identities and security groups separate from existing information security infrastructure.

Usability Testing

The goal of this pilot is to evaluate whether this data dissemination platform meets our expectations from both the technical and user perspectives. How usable Bank and System staff will find any tool is one of our primary concerns. While we were not able to engage in
any formal usability testing, we did want to get informal feedback from potential users. We worked with three volunteers to address the following questions.

- How long do users need to complete the entire submission workflow and do they consider the process burdensome?
- Do the metadata fields defined in the pilot describe the data sufficiently?
- Are users willing to provide supporting documents and further descriptions, such as a data dictionary, in the submission process?
- How effective is the user interface? Do users suggest any improvements?
- What is the role of the data curator and how could this individual help improve the efficiency of the workflow?

Each user completed the test independently to diminish peer influence. During the test, we instructed each participant to walk through the submission workflow: logging in, entering the necessary metadata, and then assembling and submitting the data file. Once the test was completed, the participant was asked to provide feedback on his or her overall impression of the tool and the effectiveness of the workflow, as well as suggestions on how to improve the user interface.

Users found some steps in the workflow challenging at first. They received registration emails for access to the platform that they weren’t sure how to handle and for which we had failed to prepare them. Even once they understood the registration process, they were somewhat stymied by a technical difficulty peculiar to this pilot: the submission process relied on two separate websites for different parts of the workflow. During the pilot, the two sites were not well integrated, resulting in some confusion as users navigated between two similar looking, but disconnected, web pages. We do not expect this problem when the platform is commercially available.

Once the users understood where to start, we observed them while they completed the two parts of the submission workflow: metadata entry and data file assembly. Metadata entry required the users to fill out all of the mandatory fields and presented optional fields as well. By limiting the number of required entries, we hoped to improve efficiency and reduce the cost to researchers for publishing their data. None of the users we observed seemed to notice the distinction between required and optional, so they simply filled out all of the blank fields.

The next step, data file assembly, required users to upload data from their computers to the Kansas City Fed staging area for the platform. As previously mentioned, users encountered some technical difficulties with the connection to the storage system, and only two of the three testers were able to successfully construct and upload a data file.

**User Feedback**

Overall, the users reported that the amount of time spent on submission was not burdensome. They also reported that a majority of the metadata fields were easy to fill out. However, the users offered a few suggestions regarding metadata and user interface at the end of the test:

- Enhance the metadata selection interface
  Users were asked to select terms from a controlled vocabulary specific to economics and finance. To encourage consistency across the Federal Reserve System, we used a list maintained by colleagues at the Federal Reserve Board of Governors to describe their research publications. The list was a flattened hierarchy of more than 70 lines and was difficult to navigate in a drop-down menu. Users suggested retaining the hierarchical structure but splitting the long list into multiple drop-downs to improve navigation and readability.

- Add detail to some fields
  The Journal of Economic Literature (JEL) maintains an alphanumeric, hierarchical classification system that is the standard for classifying scholarly literature in the field of economics. The interface for submission used the text description for the economic field without including the 2-3 digit identifier. Because economists are so familiar with the JEL scheme, users suggested adding the code designation next to the description.

- Clarify metadata fields
  The users found a few metadata fields ambiguous and suggested clarifications and examples to improve the workflow. For instance, we named one metadata field ‘Data Restriction Audience’ to contain information on time restrictions (when can the data be shared) and access restrictions (with whom can the data be shared). Users weren’t sure what the name implied, and they certainly did not understand our intended usage.

  The field labeled ‘Documentation’ also confused users. We expected users to provide supporting documentation such as a data dictionary. One economist expressed willingness to do so but was unsure how much detail was required. He also pointed out that as the definition of certain data variables has changed over the years, providing information on how the data was constructed and elaborating on the difference would add high value to data dissemination.

- Allow options for restricted data
  Many of the researchers in the Federal Reserve System are assigned to work with restricted data that can only be shared with certain audiences. The user who volunteered to test the platform using restricted data noted that some of the data could still be made available after sensitive information was extracted. Creating versions of the data that can be shared more broadly will definitely
require more work on the researchers’ side, but will eventually be beneficial to the research community within the Federal Reserve System as well as to the public.

- Ensure persistent identification

All three test users expressed concerns with the permanence of the URL for their research output. As the purpose of data dissemination is to make data available, the economists wanted to know how we would ensure consistent access to the data once published. CADRE staff were already working on a program to assign digital object identifiers (DOIs) to research output, which would ensure current location information for files and provide identifiers to published datasets.

Curation Testing

In the curation workflow, the curator reviews the submitted metadata and the data file, adds additional information as needed, and approves or rejects the submission. Though we had early success in curating entries in the pilot, we were not able to curate the data the three users submitted during this phase of testing. There appeared to be some technical issue that neither the platform developers nor the Kansas City staff could identify or resolve. After a seemingly unrelated patch application, the problem disappeared. All curation attempts thereafter were successful.

Overall, the curation workflow functioned as anticipated and worked well. We are considering changing the workflow so that the curator assembles the data file instead of the data submitter but have not yet tested this possibility.

Next Steps

To accomplish the objectives set at the beginning of the pilot, we need to take the following steps:

- Repeat the usability testing with the three economists to ensure all of them are able to submit metadata and upload data files from their computers. We will modify the curation testing to determine whether the curator is able to assemble data files for the data submitter and to ensure the submitted metadata and data are successfully curated.
- Finalize the dissemination workflow to include DOI assignments which were created locally and temporarily during the pilot. We will investigate how the publication platform integrates with our DOI registration service.
- Work with technical staff to verify the security settings of the platform. Ideally, we would implement a data dissemination platform for both public data and restricted access data. One important aspect of this pilot is evaluating the access settings of the platform to ensure they meet our security requirements.
- Expand this pilot to other interested participants such as researchers in other Federal Reserve Banks and the Federal Reserve Board of Governors. We anticipate involving more potential users and collecting feedback from them, particularly on their experiences with metadata entry, data submission, and identity authentication.

References


Notes

1. San Cannon, Federal Reserve Bank of Kansas City. Email: Sandra.Cannon@kc.frb.org.
2. Deng Pan, Federal Reserve Bank of Chicago. Email: Deng.Pan@chi.frb.org.
7. http://www.icpsr.umich.edu/icpsrweb/deposit/
IASSIST 2017

Lawrence, Kansas
May 23-26, 2017

Save the date! More information coming soon!

Links

IASSIST

Federal Reserve Bank of Kansas City
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST. Join today by filling in our online application:

http://www.iaassistdata.info/