Social Science Data Archives: Case Studies in Data Sustainability

Kalpana Shankar[1], Kristin Eschenfelder[2], Greg Downey[3], Peter Fleming[4], Cynthia Engerson[5], Rebecca Lin[2] and Jennifer Nygren McBurney[2]

[1] Dublin School of Information and Library Studies, University College

[2] School of Library and Information Studies, University of Wisconsin-Madison

[3] School of Journalism and Mass Communications and School of Library and Information Studies, University of Wisconsin-Madison

[4] National College of Ireland

[5] Kent District Library, Kent

Abstract

There has been a sizeable investment in the development of large-scale data and appropriate infrastructures in the physical and biological sciences and increasingly in the social sciences and humanities. Concerns about data sustainability have attracted a great deal of attention as research project data collection represents a significant investment, and loss of subsequent use of that data represents a loss of potential value. In this poster, we focus on of the most long-lived examples of data archives: Social Science Data Archives (SSDAs). In this study, we report on preliminary research on the historical, institutional, and operational dimensions over SSDAs over time. Drawing upon analyses of institutional and policy documents and interviews with staff, depositors, and administrators, this poster briefly discusses current challenges to SSDA longevity and implications for next steps in expanding the study both theoretically and methodologically. Initial themes discussed in this poster include data archives making a market for themselves, configuring their products and their user base and ongoing tensions between the need to generate revenue and pressure for open access data.

Keywords: social science data archives, data archives, case studies, data sustainability, history, information infrastructure

Citation: Shankar, K., Eschenfelder, K., Downey, G., Fleming, P., Engerson, C., Lin, R., & McBurney, J. N. (2014). Social Science Data Archives: Case Studies in Data Sustainability. In *iConference 2014 Proceedings* (p. 864–868). doi:10.9776/14287

Acknowledgements: Funded in part by a grant from the University of Wisconsin Alumni Research Foundation and the Irish Research Council New Foundations Scheme.

Contact: kalpana.shankar@ucd.ie, eschenfelder@wisc.edu, gdowney@wisc.edu, peter.fleming@alumni.nuim.ie, Cynthia.engerson@ucd.ie, lin.rebecca.k@gmail.com, nygren.jennifer@gmail.com

1 Data Archives and Sustainability

This project examines strategies employed by data archives to remain sustainable over time. It focuses on the most long-lived examples of data archives: Social Science Data Archives (SSDAs). The social sciences have enjoyed stable data archives since the 1940s (Green and Gutmann, 2007). Their longevity provides an opportunity to examine archive sustainability through changes in funding, technology, and organizational infrastructure.

We report on one historical case study: the Inter-university Consortium for Political and Social Research (ICPSR). We compare the historical themes to explore a much smaller and newer SSDA, the Irish Social Science Data Archives (ISSDA).

We explore two research questions:

- What are some key strategies employed by ICPSR to remain sustainable?
- How do the historical themes from ICPSR compare with contemporary issues at ISSDA?

2 The Emergence of SSDA

Approximately 16,000 PhDs entered the social sciences in the two decades following the Second World War (Lefcowitz and O'Shea, 1963). Emphasis on quantitative research in the social sciences — especially in political science, mass communication and economics, but also in history, sociology, and anthropology — grew as more public opinion surveys, economic development indicators and geocoded election data become available, as well as the computer hardware and software to generate and process them. Data archives such as the Roper Center began acquiring machine-readable data in the 1940s and first opened to the general public in the 1950s (Scheuch, 2003). In both Europe and the US, the 1960s saw the establishment of over a dozen data archives, consortia, and dedicated library services to coordinate data collection efforts across institutions, to promote sharing of valuable data and to educate students and scholars about quantitative and machine processing analysis methods (White, 1977) . New professional associations such as the International Association for Social Science Information Services and Technology (IASSIST) in the US, the Federation of European Social Data Archives in Europe (FESDA) and the International Federation of Data Organizations for Social Science (IFDO) emerged to help develop professionals to manage social science data sets and their processing technologies and coordinate international standards efforts and data exchange (Nasatir 1973).

3 SSDA Cases and Methodology

ICPSR, located at the University of Michigan in the United States, is a "consortium" composed of members who pay dues for access to data and have voting rights within the ICPSR governance structure. ISSDA, based in Dublin, Ireland is not a membership organization, does not have dues, and is primarily funded by the University College Dublin library budget (UC Dublin is a publicly funded university). Both curate data (as opposed to simply hosting raw data), offer value added services such as online analysis tools, and support educational services or materials.

For our institution level analysis, we examined 40 years of ICPSR records, including thrice-annual governance meeting minutes, annual reports, strategic plans and grant proposals. We conducted semi-structured interviews with current and former staff and with researchers who have participated in governance. At ISSDA, we conducted interviews with institutional managers, depositors, and users of ISSDA. For our field level analysis, we examined documentation and conference proceedings from data professional organizations from the 1970s to the present including the IFDO, CESSDA, and IASSIST.

Analysis of the data is inductive, iterative and informed by the investigators' theoretical orientations. We first read over all historical documentation, making notes and tags in Mendeley. We then organized key themes over time. We summarized conference proceedings in a similar fashion. We compared institutional-level ICPSR themes with field perspectives provided by the professional society documents. Analysis is ongoing; this paper reports preliminary results. We also compared ICPSR themes with findings from the contemporary ISSDA case study.

4 Initial Findings

In this poster we present three preliminary themes from our historical analysis of the ICSPR data: making a market, product and user configurations, and open data and the commons.

4.1 Making a Market

ICPSR took various actions over time to shape social science research to align with the ICPSR mission. It created a need for itself through the popularization of quantitative computer-assisted analysis of ICPSR data sets. ICPSR's "summer program" workshop series trained researchers and students and mobilized them as advocates for ICPSR. To use newly acquired skills, researchers needed access to large sets of machine-readable data, which were not readily available outside ICPSR. ICPSR fashioned itself as an essential

service of social science research by soliciting data from researchers and government agencies for its members, curating the data by cleaning it, de-identifying it and providing documentation, and then facilitating access and use with users at member universities through a system of "member representatives" or data library professionals. Member representatives would take requests from end users and receive the data from ICPSR, first through punch cards or tape drives, then later CD-ROMs. ICPSR cultivated relationships with member representatives, so that they would also advocate for ICPSR membership within their institutions.

ICPSR also sought to coordinate collections with other data archives to avoid creating competition for data sources. For example in the late 1960's ICPSR and Roper (also a membership data archive) coordinated which archive would collect and distribute which data. ICPSR also ensured its longevity through standards and professionalization of data practices, and positioning itself as a source of new knowledge about how to archive data. For example, in the mid 1990's ICPSR convened the first committee to look into standards for social science data (i.e., DDI). The summer program incorporated classes on data curation, and ICPSR changed staff position descriptions to encourage more research activity on curation.

4.2 Product and User Configuration:

Throughout its history ICPSR has made significant changes around users, the nature of its products, and its pricing.

In the early days, ICPSR sought membership and dues primarily from political science and sociology departments at large research universities in the US. As time went on, ICPSR re-positioned itself as something libraries should purchase because individual academic departments could no longer afford ICPSR fees. Over time with the large US research university market tapped, ICPSR sought to increase international and smaller college membership. Along the way, ICPSR also allowed (non-voting) membership by government agencies and commercial organizations.

ICPSR has sought funding from many different governmental and foundation sources over time. For example, their early training funding from the National Science Foundation restricted them to disciplines NSF considered "scientific" (i.e., few historians), so ICPSR instead appealed to IBM for historian training funding. ICPSR has received significant funding from the US Department of Justice (over 14 million), National Institutes of Health (over 8 million) and and Health and Human Services and the National Science Foundation (over 5 and 4 million respectively). Foundation sources like the Robert Wood Johnson Foundation and Mellon have provided more modest inputs.

ICPSR originally shipped punch cards, tape drives and then CD-ROMs of requested data sets to organizational representatives (ORs) at members' universities. These ORs would then facilitate access for end users and store the data for local reuse. In the late 1990s, in a major change in distribution patterns, ICPSR provided online access directly to some end users via a service "ICPSR Direct" that relied on new networking infrastructures. The nature of the ICPSR product changed based on perceived data demands within the scholarly fields; for example, scholarly trends encouraged diversification from US voting and political science data to acquisition of more international and economics data.

Most importantly, ICPSR began to offer data hosting services to government agencies in the 1970s. Grants and contracts to host "topical archives" for agencies became the largest source of revenue for ICPSR after 2002, such that by 2010 it was just under three times ICPSR's membership income.

ICPSR adjusted its pricing options over time to either attract new desired customers or deal with access problems created by technologies like computer networking. ICPSR's primary pricing model included several different "classes" of universities reflecting differences in whether a university had graduate research-oriented programs, or simply undergraduate programs. A further challenge emerged in the 1970s with computer networking as member campuses began to share access to ICPSR data with nonmember institutions via regional networks. ICPSR responded by creating a new "federated" pricing model that

accommodated low volume access to new campuses via computer networking with member institutions. It also added an international membership category. In 1997 it offered membership to the OhioLink consortia in the US. ICPSR has always allowed access to particular data sets for one-time fees.

4.3 Open Data and the Commons

From its inception, ICPSR experienced a tension between the archive's mission to encourage as much use as possible and the need to generate income in order to process and curate data and offer educational services. ICPSR data collections had typically only been available to members. But, the ICPSR Board has always considered and granted "free" access to data in instances where they perceived the scholar's home institution might not be able to afford a subscription or purchase of data sets. Freeriding by economically viable institutions whom the Board perceived ought to be members was not permitted.

Pressure for "free" access also came from other institutions that began to provide free public access to government data and from the government agencies which paid ICPSR to host their data. For example, University of Minnesota hosted the 2000 US Census data for free, drawing ICPSR's members only access into question. By the late 1990s the Membership committee reported that some schools were dropping membership because the data faculty needed were available for free.

Some government agencies who paid ICPSR to host data also required free public access to some data. In 1998, about 20% of ICPSR's data was available for free from its website including the Substance Abuse and Mental Health archive hosted for the US Department of Health and Human Services, and the National Archive for Computerized Data on Aging hosted for the US National Institute on Aging.

4.4 Insight into Current Issues: ISSDA

One challenge that ICPSR has had to address over the course of its history is making the case for itself to funding agencies and researchers, something that is only happening in the twelfth year of ISSDA's existence (during a time of extreme economic austerity). While ICPSR has not solved the problem of revenue, it has adapted by creating new products and customers. ISSDA's current challenges stem from a combination of historical and contemporary factors: a complicated history of staffing and funding, uncertain ongoing funding, and lack of specific expertise in working with social science data. While it has not created new products, since research began on this article, ISSDA has taken other steps towards sustaining itself. These steps include affiliation with ICPSR and membership in the European Research Infrastructure Consortium (ERIC) to increase visibility and gain SSDA expertise. ISSDA is also increasing outreach to potential depositors (particularly large grant recipients) and implementing a streamlined process for data deposit and use (thus potentially increasing both).

5 Conclusion and Next Steps

Future research will include more cases to obtain a fuller picture of SSDAs over time. Furthermore, our initial studies suggest that analyzing SSDA as "knowledge commons" (Hess and Ostrom, 2007) will enrich our understanding of SSDAs as institutions and sets of practices. Subsequent work will include more cases of SSDAs that serve different fields with different changes in science data practices. Our study of the history of SSDAs in curating data and maintaining access will provide insight into the sustainability issues of data archives in other fields.

6 References

Green, Ann, G., Gutmann, Myron, P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. OCLC Systems & Services, 23(1):35–53.

Hess, C.; Ostrom, E. (2007) Understanding Knowledge as a Commons: From Theory to Practice. Cambridge, MA: Oxford University Press.

Lefcowitz, M., and O'Shea, R. (1963). A proposal to establish: A national archives for social science survey data. American Behavioral Scientist. Retrieved from http://abs.sagepub.com/content/6/7/27.full.pdf+html

Nasatir, David. (1973). Data archives for the social sciences: Purposes, operations and problems. Paris: UNESCO.

Scheuch, Erwin K. (2003). History and visions in the development of data services for the social sciences. International Social Science Journal 177: 385-399.

White, Howard D., ed. (1977). Reader in machine-readable social data. Englewood, CO: Information Handling Services.