“The Price of Keeping Knowledge” Workshop: ICPSR Position Paper

Jared Lyle
George Alter
Mary Vardigan

Overview

The challenge of long-term funding for institutions providing public goods is not unique to digital preservation. Communities have a variety of ways of supporting memory institutions, like libraries, over the very long term. Long-term access to data requires durable institutions that plan on a scale of decades and even generations. Such planning is difficult when grant cycles are of limited duration, and proposed projects are rated for innovation and transformation but not for reliability or permanence. Fee-for-services can provide additional financial stability, although it can be difficult to predict accurate and comprehensive costs based on future needs and expectations.

The Inter-university Consortium for Political and Social Research (ICPSR), a center in the Institute for Social Research at the University of Michigan, has a 50-year record of archiving and disseminating social science data.[1] ICPSR is committed to ensuring long-term access to the more than 500,000 files that comprise our 8,000-plus research collections. As the demands of the scientific community increase, ICPSR is exploring new ways to sustain access to its collection for future researchers.

ICPSR’s Current Model

ICPSR has developed a sustainable funding model based on a large user base and diversified funding sources: membership subscriptions from over 700 institutions for data users, and dissemination services paid by grants and contracts from twenty different Federal and private funding agencies. For the 2011-12 fiscal year, membership dues income contributed 18 percent of ICPSR’s total revenue, while grants and contracts accounted for 51 percent. Additional revenue came from Summer Program tuition and other sources.[2] Funding for ICPSR is not underwritten by the University of Michigan, and we do not have long-term commitments from any of our sponsors.

This diversified revenue pays for curation services to prepare and describe data, preservation services to store and migrate data, and dissemination services to distribute data. ICPSR does not itemize per-dataset costs; rather, each sponsored archive within the organization prioritizes how many data collections to archive within a fiscal year. These priorities are usually guided by the funder. One archive may devote extensive resources to intensively curate, preserve, and

[1] http://www.icpsr.umich.edu

[2] Complete financial details are provided in the 2011-2012 ICPSR Annual Report, here: http://www.icpsr.umich.edu/files/membership/or/annualreport/2011-2012.pdf

disseminate just a handful of data collections, while another archive with similar funding might choose to preserve and distribute hundreds of files, albeit with lower levels of curation.

New Models

As the research landscape has changed, new challenges have emerged. First, the cost of protecting confidential information about research subjects has been rising. Research designs in the social sciences are more likely to include elements that make it easier to identify human subjects, such as longitudinal data, geospatial locations, and multi-level data (e.g. student, teacher, school). Analyzing data for disclosure risks has become more complex, and measures to mitigate those risks are costly. When data cannot be modified to minimize the risk of deductive disclosure, ICPSR provides access under data use agreements and other procedures to assure that subjects are protected. Again, these measures are costly. Consequently, the minority of researchers who use confidential data generate a disproportionate share of the costs of data archiving and dissemination.

Second, social science researchers are using new kinds of data, which pose new problems and higher costs for long-term preservation. ICPSR recently acquired a large collection of videos for the Measures of Effective Teaching (MET) Project funded by the Bill and Melinda Gates Foundation. Since video files are much larger than files containing quantitative data, the MET video collection is larger than the sum of all of the files archived at ICPSR in its previous fifty years. Even larger collections may arrive in the near future. Internet data (transactions, clicks, tweets, etc.), sensor data, and other new types of data are emerging every day.

Third, funding agencies are moving in the direction of open access requirements for data. While we applaud the motivations behind the open access movement, the implications of open access mandates have not been fully examined. Who will bear the costs of documenting and preserving all of these data collections? How can limited resources for data archiving be focused on data with the highest value for secondary analysis?

We are investigating a number of models, which are described below. Since we believe in the importance of diversification, these would enhance, not supplant, our current model.

a. *Fee-for-services model*. As the costs of services provided to data users have become more differentiated, we are exploring fees for specialized services. Fees would offset costs for resource-intensive collections and services. For example, users who analyze confidential data through a secure remote access facility may be charged for this service. Since these fees may pose financial challenges to some users, there may be preferential rates for students or other groups.

b. *Infrastructure model*. Some research councils in Europe are considering long term commitments to data archives as necessary scientific infrastructure. We advocated for this model in a recent response to the White House request for information on open access to data, stating: “A Federal program to establish and support long-lived institutions is needed to create

repositories capable of providing preservation."[3] A long-term commitment would assure that an archive would persist beyond the current three- to five-year grant cycles, but periodic competitions among service providers are also possible. Funding agencies in the U.S. are very reluctant to make long term commitments of this kind.

c. *Endowment model*. Long term data archiving could be funded by providing an endowment for every data collection. The future costs of preservation and dissemination would be captured when a data collection is deposited in an archive. The endowment would be estimated to cover the present discounted value of all future costs associated with data curation, dissemination and preservation. Endowment funds would be invested and drawn down as needed. The endowments created in this way would provide a buffer against future uncertainties, help articulate preservation costs for the archive, and indicate real archival costs for depositors. This model poses some challenges, however. In the U.S., federal agencies cannot allocate funds for future costs in grants and contracts. It would be even more problematic if these agencies tried to calculate these costs when a new project is funded, because it is very difficult to anticipate what data will be produced by a new project.

d. *Re-charge model*. In this model, charges paid by new data collections cover the costs of maintaining earlier deposits. Since current operating costs are known with reasonable certainty, it is easier to estimate this model than the endowment model. Depositors must be convinced that the future benefits that they receive justify their support for earlier data collections. This model may require either rising fees or increasing flows of data, because the costs of servicing the entire collection are likely to rise as the archive grows.

Summary
ICPSR's current funding model built on a large user base and diversified funding sources has sustained curation, preservation, and dissemination efforts for 50 years. That said, we are continually reviewing the robustness and diversity of our funding and recognize the need for ongoing innovation.

We are encouraged by and monitor recent projects and dialog within the preservation community about financial models.[4] Indeed, we recently received a Sloan grant to explore a variety of issues related to open access to data, including the sustainability of data repositories.[5] Making explicit our costs is something toward which we strive, especially with respect to preservation costs, which are often conflated with other expenses, leading to potential underfunding and underappreciation of the time and effort required to protect valuable data resources.

---

[3] http://www.whitehouse.gov/sites/default/files/microsites/ostp/digital-data-(%23043)%20ICPSR%20Response.pdf

[4] We have noted the many recent forums for discussing preservation and curation costs. See, for instance, here: http://wiki.opf-labs.org/display/CDP/Home

[5] http://www.ns.umich.edu/new/releases/20812-u-m-sloan-project-enhances-open-access-to-research-data