The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons

Limor Peer,

Yale University

Ann Green,

Digital Life Cycle Research and Consulting

Abstract

In 2009, the Institution for Social and Policy Studies (ISPS) at Yale University began building an open access digital collection of social science experimental data, metadata, and associated files produced by ISPS researchers. The digital repository was created to support the replication of research findings and to enable further data analysis and instruction. Content is submitted to a rigorous process of quality assessment and normalization, including transformation of statistical code into R, an open source statistical software. Other requirements included: (a) that the repository be integrated with the current database of publications and projects publicly available on the ISPS website; (b) that it offered open access to datasets, documentation, and statistical software program files; (c) that it utilized persistent linking services and redundant storage provided within the Yale Digital Commons infrastructure; and (d) that it operated in accordance with the prevailing standards of the digital preservation community. In partnership with Yale's Office of Digital Assets and Infrastructure (ODAI), the ISPS Data Archive was launched in the fall of 2010. We describe the process of creating the repository, discuss prospects for similar projects in the future, and explain how this specialized repository fits into the larger digital landscape at Yale.

Background

The Institution for Social and Policy Studies (ISPS) was established in 1968 by the Yale Corporation as an interdisciplinary center at the university to facilitate research in the social sciences and public policy arenas. In 2001, ISPS announced the Experimental Initiative, designed to encourage field experimentation in the social sciences at Yale. The term 'field experiment' refers to fully randomized research designs (often called Randomized Controlled Trials, or RCTs) in which observations found in a naturalistic setting – voters, patients, welfare recipients, community organizations, government entities, and the like – are assigned to treatment and control conditions. Recent examples of this kind of research at ISPS include randomized studies of voter mobilization, peer counselling of homeless people, campaign activities in Africa, and the persuasiveness of televised campaign advertisements.[1]

Many researchers affiliated with ISPS incorporate field or other experiments (i.e., survey, natural, lab) in their research design, and as such comprise a specialized research community. Over the past fifteen years, this community has grown as graduate students and post-doctoral fellows leave Yale to join other institutions. In Political Science, the significance of experimental design and methods has also grown, as evidenced most recently by the establishment of a new section of the American Political Science Association in 2010 named "Experimental Research."

The ISPS Data Archive is a digital repository for research produced by scholars affiliated with ISPS, with special focus on experimental design and methods. It launched in September 2010. The Data Archive is meant to capture and preserve the intellectual output of a single unit within the university, and to serve as a model for preserving research data.[2] ISPS began developing the Data Archive as part of an overhaul of its research management practices that included better tracking of research from conceptualization, through research design, funding and IRB approval, and on to publication. The first step in these efforts was to revamp the ISPS website, as shown in Figure 1.Utilizing WordPress blogging and content management software, the daily management of web content was improved. More importantly, it offered a technical solution for both better management of the research process itself (e.g., online grant application forms alerting research staff to potential new research projects are now accessible through the ISPS website), and a means for making the research more open and accessible.

The revamped ISPS website was designed to integrate key research information and documentation – Projects, Publications, and Data – into a common interface. In August 2009, two of the three elements of the new Research section of the ISPS website, Projects and Publications, were introduced at the launch of the site. The Projects database includes planning information communicated before the research is carried out, including research design and analysis plans, budget, timeline and sample

---

[1] For further examples, see the ISPS website at: http://isps.research.yale.edu/

[2] While there are data repositories in the social sciences, most notably the Inter-university Consortium of Political and Social Research (http://www.icpsr.umich.edu/icpsrweb/ICPSR) at the University of Michigan, there are currently no specialized repositories for experimental data. MyExperiment (http://www.myexperiment.org), an open access and collaborative space, focuses on scientific workflows.

size, and tracks status changes over time. The Publication database includes basic citation information and links to published journal articles. We then turned our attention to the third element: Data. At this point, ISPS partnered with Yale's Office of Digital Assets and Infrastructure (ODAI) to find solutions relating to storage, persistent linking, long-term preservation, and integration with a developing institutional repository. As such, ISPS became a pilot project within the full institutional context, helping to identify issues specific to digital preservation of data across the campus.

Figure 1. ISPS website.

Guiding Principles

The ISPS Data Archive has four inter-related guiding principles: replication, integration, open access, and stewardship.

Replication

The Data Archive was primarily created in order to support a service model of managing data with a view toward usability and replication. Replication, defined as "the confirmation of results and conclusions from one study obtained independently in another—is considered the scientific gold standard" (Jasny et al., 2011), is a key principle in the development of the Archive. The Archive is meant to be used for reproducing the experimental results through replication, i.e., by using author-provided code and data (see Stodden, 2009). Accordingly, the ISPS Data Archive provides three categories of information necessary for replication of experimental results for every study: the raw data, metadata, and the statistical code that produced the original results. In particular, sharing statistical code is key to enabling replication (Anderson et al., 2008; Peng, 2009, 2011; Stodden, 2011).

The specialized nature of experimental data requires high quality documentation and metadata to facilitate replication of, and provide meaning to, the study to enable sharing and to aid in data discovery. The production of this metadata is labor-intensive

and requires dedicated staff and dollars. As Gurstein (2011) explains, with regards to a survey successfully used by political authorities in California to make health policy decisions, attention must be paid and resources must be provided "to ensure that the data is usable by those who might make effective use of it." Our experience teaches us that the goal of publishing replicable data prescribes that steps be taken to ensure that the relevant materials are shared and that they can be used in the long run. That is, that users will be able to easily, freely, and independently make use of these materials and, perhaps more importantly, that they will be useful to them.

Integration

The ISPS Data Archive places a premium on integration in three different spheres: integration with other elements of the ISPS website (Projects and Publications), integration within the research process, and integration with elements outside ISPS (interoperability). The first type of integration means that ISPS is committed to offering users a single-point of access to research materials from its website, including the research plan and registration of the study ("Project"), the published results ("Publications"), and now the actual data collected and the statistical code used to analyze it ("Data") as shown in Figure 2. Linking publications, planning materials and data is good practice and follows directly from the first principle, Replication:

> "A journal article describing the results of scientific work is typically a distillation of experimental data aimed at a wider audience than the immediate peers of the authors. Generally inferences are made only from the most pertinent results, which are reported in a summary format, and journal publication is detached from the production of the experimental data. This renders replication or reuse of the data impossible and results in severe information loss." (Coles, Frey & Carr, 2007; see also Bourne, 2011)

Figure 2. ISPS research process integration.

The second type of integration focuses on the researcher's processing needs and provides methods of tracking research from its inception to its conclusion. The goal is to make preserving and sharing research materials an integral part of the research process. By integrating our Projects, Publications and Data on the website, the system is designed to support as much transparency as possible for the benefit of the community, and as much workflow support as possible for the benefit of the researcher. As a comment on a JISC listserv on repositories recently stated, an ideal repository is one that: "...accepts deposits from me as part of *my* work processes invisibly without any additional work, or even having to think about it" (Franklin, 2011).

The third type of integration – how the ISPS Data Archive fits into the infrastructure at Yale and beyond – raises issues of interoperability for discovery and access, and is discussed below.

Open Access

A third guiding principle for the ISPS Data Archive is open access. The Archive strives to provide free and public access to research materials in line with open access principles (e.g., OECD Declaration on Access to Research Data from Public Funding, (2004); Budapest Open Access Initiative by the Open Society Institute, 2002) and in step with growing user expectations about the availability, usability, and quality of research data. Access to files allows members of the scholarly community to validate the existence of a specific set of data, to acquire such data (when permitted), to replicate analyses, and to view additional materials associated with a given study, including high quality metadata. The open access principles are closely linked with the principle of replication because access is necessary for others to reuse the materials and replicate the results. In practical terms, the Data Archive is set up to make data accessible directly, not through gatekeepers such as publishers and foundations, and at no cost to the user. Also, to facilitate discovery and use, materials are presented and organized from the users' perspective (i.e., by study) and are searchable within the site. ISPS is committed to making data available to researchers while taking into account the legal framework of intellectual property rights and privacy regulations. In rare cases, users are required to contact the original researcher for permission to use confidential data or meet certain conditions of use. However, ISPS seeks to make these restrictions the exception rather than the norm.

Stewardship

The goal of the Data Archive is to preserve and provide ongoing, persistent access to a body of knowledge generated by a specialized community, including statistical data. The care and effort put into enabling replication, integration and open access would be futile if the digital materials are not kept safe for the long term. The goal is to always add data – never deleting it – and to maintain its accessibility and usability by converting files to readable formats and open source programs (e.g., text, ASCII, R) as formats change over time. This long-term view requires a commitment to the ISPS Data Archive, on behalf of ISPS and Yale, both in technical and financial terms.

We now turn to a description of the features of the Data Archive and how it works, followed by some of the challenges yet ahead.

Features

Content

The ISPS Data Archive includes a growing collection of research output of ISPS-affiliated scholars for the purpose of replication. This research output – including data, metadata, statistical code, codebooks, research materials, and description files – originates primarily from field experiments in the social sciences. This is original, often "small" data with high value for a growing community of researchers, educators, policy makers and students (Palmer et al., 2007). The goal is for each set of of files to be self documenting and sufficient for the replication of each study's results. Data files are primarily "born" digital, but may come from survey data or hand-entered data, and they utilize a variety of software, as do the statistical code files. Metadata documentation is created at the study level, the file level, and the variable level, and is Dublin Core- and DDI-compliant. The files are organized by published article, and as such they are linked to the Project and Publication content for the same study via the ISPS website. The collection is small, currently consisting of over 40 studies and a total of about 700 files, taking up about three gigabytes. Data deposits are not mandated, but encouraged. New studies make up the majority of the deposits, but there is also a backfill of older studies.

Process

The ISPS Data Archive is managed by a full time professional with knowledge of the specific research domain. A team of graduate student research assistants is recruited to handle most of the file processing. Yale's Social Science Statistical Laboratory (StatLab) provides assistance in converting proprietary statistical code to R, which is open source. Currently, file handling and processing is minimally automated. In broad terms, the process of data includes the following steps: convert files to user-friendly formats, update these formats as appropriate, maintain permanent backups of the digital content of the replication files, review files to determine whether any issues of confidentiality exist, further review program files and create identical files in R, prepare metadata records, including searchable fields, to assist in locating files within the ISPS Data Archive, and publicly announce the availability of data on the ISPS website and elsewhere. The most labor intensive steps in the process include the data and code checks, such as checking for sensitive data, completing missing variable information and checking the statistical code, as these are closely scrutinized. Quality checks are conducted throughout the process via regular communication among team members.

Note that the quality control and standardization of files is beyond what most researchers will do (McCullough, 2007; McCullough & McKitrick, 2009). The strength of this archive is that it provides its researchers value and reduces the cost and effort of sharing data meaningfully. The incentives for researchers to take part in the ISPS Data Archive stem from their ongoing relationship with ISPS as a funding source and include fact-checking prior to publication, seamless use of research assistants by the Archive, and the opportunity to archive proprietary data until such time as it can be released. On their part, researchers are asked to clarify rights to and ownership of the data, and to ensure that they are not in breach of any laws or

contracts. They are also asked to remove personal identifiers contained in files that could allow direct or indirect identification of individuals.

Infrastructure

The Research section of the ISPS website, which contains the ISPS Data Archive, was custom-designed and programmed by an outside vendor using WordPress. This vendor continues to maintain and support the website. Managing the MySQL database, as well as the data management and archival activities, is a specialized and labor-intensive task, requiring a designated person at ISPS. ISPS benefits from close consultations with the University's social science librarians and data specialists regarding protocols, standards and software. The ISPS Data Archive also greatly benefits from Yale's institution-wide services. Yale's Office of Digital Assets and Infrastructure (ODAI) provided support with planning, strategy and implementation. ODAI also enabled the use of the Yale Digital Commons[3] replicated disk storage to host the Data Archive in a stable and secure environment, and the Yale Persistent Linking Service for handles-based linking to each file.

Users

Users of the ISPS Data Archive are primarily social scientists from a variety of disciplines. This community uses state-of-the-art statistical methods and programs. It has high expectations about availability, usability and quality of data and metadata. As this community is increasingly inter-disciplinary, care is given to creating an environment that is not discipline-specific. While the community is small, it is growing both within Yale and beyond.

Access

Generally, the Archive provides 24/7 access to the website and data files without restrictions or required permission, although users do have to consent to the ISPS Terms of Use. As the public files in the Archive are intended for replication, ISPS is licensing them under Creative Commons.[4] A small number of files are restricted due to confidentiality or other considerations, with an option to contact the authors. Users download data and import to statistical programs (e.g., SPSS, Stata, R) on their own workstations.

Early Feedback

Some evidence of success comes from the ISPS community itself: the ISPS Data Archive inspired a change in thinking about research data management and the need to have a data plan. Having gone through the experience of digging up old files for deposit and enduring back-and-forth communication with the Data Archive team to clarify information about the files, researchers now plan ahead and submit much cleaner and more complete sets of files. They also appreciate that they can point to the infrastructure already in place when writing their data management plans, that their

---

[3] Yale Digital Commons is a collaborative framework for developing services to support the lifecycle management and use of Yale's digital assets. See: http://odai.yale.edu/projects/yale-digital-commons

[4] Creative Commons: http://creativecommons.org/

research output will always "be there," and that the Data Archive provides a comprehensive, single-point access to all ISPS materials. An important benefit to researchers is the data and code checks by Archive staff, which occasionally turns up minor errors that can be fixed before publication. Other benefits, mentioned to us anecdotally, include pride and a sense of community as a result of clearly seeing the inventory of ISPS research. Researchers also expressed appreciation for the alerts that are sent to the wider community when data become available through the Archive, leading to a reduction in the number of direct requests for data going to researchers. As this is early in the life of the ISPS Data Archive, much of the focus is on building trust (Steinhart et al., 2009) and cultivating a community of users who will use the Archive as a high-quality and reliable resource. Conversations with our researchers reveal that there are things they value which the ISPS Data Archive is not currently set up to provide, including a staging space for in-progress work, easy-deposit connections to their regular working environments, and version control.

On the demand side, we have some data on user behavior. According to Google Analytics, the Data section on the website is the second most common entrance to the site, after the home page, and it receives the third highest number of unique page views, after the home page and main programs page. Overall, traffic is stable (about 100 page views per month), with some high points corresponding to community announcements. In addition, the Data section of the site consistently has a higher page-per-visit ratio than the site in general. Other measures might prove instructive (e.g., number of depositors divided by total items deposited; number of users by total number of views and downloads; citations and in-links), but it is too early in the life of the Archive to see trends. As some in the data preservation community have pointed out, there is a difference between use and usefulness. Anecdotally, we know that the Data Archive has some use for the broader community, as we have received requests to register experimental studies from scholars not affiliated with ISPS. In addition, published articles re-using the Archive's data have begun to appear (see for example, Panagopoulos, 2011) and the Archive has received private requests for grouping studies to aid in teaching. These are indications that the ISPS Data Archive is perceived as a valuable destination and resource. Otherwise, we hope that users will come to appreciate the convenience of a one-stop-shop, the transparency of the research process, and the educational potential of the content.

Challenges

The idea for this particular data archive started as a simple desire to share research outputs and enable the replication of research. However, it soon became clear that committing to this project requires that resources be allocated and plans be made, both at the unit and the university level, for long-term stewardship of the content in this archive, as well as more broadly for research data from across the university. In the process of developing and implementing the ISPS Data Archive, we have encountered challenges in several areas: policy, technology, sustainability, extensibility and scalability, interoperability and buy-in. It is crucial to engage with these challenges in order to provide stewardship for the Archive and to support the Archive's other guiding principles. These challenges require comprehensive responses, as they cannot be solved by one academic unit alone. Policy challenges, in particular, need to be worked through within the broader context of the institution.

The specific policy challenges that the ISPS Data Archive faces are ownership, intellectual property rights, and data protection. ISPS takes these considerations very seriously while working to create a policy framework that will support the principles of replication and open access that guide the Archive. With respect to ownership, most data are collected by ISPS researchers and may also include administrative or survey data collected elsewhere. Resolving the issues of ownership, rights, and data protection is crucial if these materials are to be shared. With guidance from Yale's General Counsel's Office and ODAI support, ISPS has drafted a deposit agreement which asks authors to identify authorship, to comply with applicable copyright laws, and to certify that they have used due diligence to prevent disclosure of the identities of research subjects. Files that contain sensitive data are restricted, but future policies might allow tiered access. On the user side, we require agreement to our terms of use policies before downloading any file.[5] These policies are in development, as well as policies on issues such as retention and deletion of data. ISPS is working with an institution-wide group sponsored by the ODAI to address policies, procedures, and plans for access, stewardship and preservation, not just for ISPS but for the university as a whole.

A second challenge is technology. Relying on WordPress for the web interface of the ISPS Data Archive has several benefits, most importantly the ability to integrate it with other elements on the website, but it also has some limitations. In our case, we rely on an outside vendor to update the programming and to synchronize it with newer versions of WordPress. Is that the optimal model? Or is in-house tech support preferable? This relates to policy questions about the larger institution's long-term commitment to these type of activities. For the university to provide this level of service, there needs to be a commitment in terms of university resources, including hardware as software as well as the implementation of policy. Another technological challenge has been simplifying and automating the process, including file transfer, handle assignment, and linking a collaborative space to the storage space.

Sustainability is a third challenge. As a pilot, ISPS benefits from the university's support in terms of storing digital content within the Yale Digital Commons, but it's not entirely clear who will bear the costs of stewardship in the long run. These long-term preservation costs include managing the content, sustaining high quality file processing and metadata, and complying with current and future regulations. Closely related to sustainability are questions of extensibility and scalability: how can the ISPS Data Archive sustain the kind of on-the-ground local support it now has at the unit and the university level? Can it be managed in a more automated way, or is hands-on domain knowledge a requirement for success? How can the ISPS Data Archive framework be extended to other types of data or research paradigms? Or to other types of uses, such as teaching and learning? How can it be scaled up to larger centers, with more researchers and more research output?

Another challenge is the issue of data and metadata interoperability. The ISPS Data Archive strives to communicate with other existing structures (e.g., ICPSR) and to optimize search and discovery. It does so by adhering to prevailing metadata standards, including OAI-PMH, Dublin Core, and the Data Documentation Initiative.[6] While interoperability may be more easily achieved with off-the-shelf digital

---

[5] ISPS Terms of Use: http://isps.research.yale.edu/research-2/data/terms-of-use/

[6] Data Documentation Initiative: http://www.ddialliance.org/

repository solutions or by including the ISPS content in larger repositories, creating a solution that balances the goal of interoperability for search and discovery with the specific needs of ISPS is tricky, as we're charting new waters. For example, an institutional repository may well simplify the deposit processes and satisfy long term storage needs, and it will certainly help with searchability across Yale's digital resources, but it may be limited with regard to the website customization and integration that ISPS is committed to. Moreover, it will not provide the type of on-the-ground support for preparing and managing research output that ISPS offers.

Finally, we should mention the challenge of buy-in from the ISPS community. ISPS has the advantage of a small, motivated group with an incentive to get the word out about its scholarly accomplishments, so participation has not been a huge hurdle for ISPS. However, other institutions have identified the difficulties of researcher cooperation with open archives, especially in light of other outlets (e.g., personal websites), requirements (e.g., journal websites), considerations (e.g., fear of being scooped, fear of revealing sensitive information), and incentives (e.g., little reward for sharing data) (Banks, 2011). ISPS strives to be relevant and provide additional services to researchers who have, over time, developed independent practices and may have chosen to distribute their work through other methods. This is especially important in light of increasing requirements for data management by funders, such as the National Science Foundation (NSF) in the U.S.

Conclusion

To sum up, the ISPS Data Archive of social science experimental data benefits from the support of the university's technological, policy, and infrastructure framework being developed through ODAI. The Archive is guided by the principles of replication, integration, open access, and stewardship. The content of the Archive consists of data and code files and documentation which allow scholars to replicate experimental social science research. The staff supporting the Archive accept the original deposit and prepare data and documentation files for dissemination with metadata creation, statistical code check, sensitive data check, file normalization, archival, and web upload. Researchers, who are increasingly required by funding agencies and journals to provide data management plans and make their data available, benefit from the Archive's services. The Archive is a valuable tool for raising awareness about data management and archival, and for educating researchers about the importance of data management earlier in the research process and throughout the data lifecycle. Finally, the Archive's focus on replication of research findings adds value to the files themselves by submitting them to a quality review and by preparing them for long term usability.

Plans for the future include pressing on with growing the collection, including developing subsets of datasets for specific uses (e.g., teaching) and engaging with the ISPS community through communication, participation, and better usage statistics. At the same time, ISPS will continue the development of policies around ownership, rights, and data protection; finding ways to create technological solutions to the needs of the Archive, including automating and simplifying the deposit and processing of files, refining access, and linking to cyberinfrastucture within and beyond the institution; sorting out issues of sustainability and stewardship at the unit and university levels; and planning for extensibility, scalability and interoperability.

Acknowledgements

References

Anderson, R.G., Greene, W.H., McCullough, B.D., & Vinod, H.D. (2008). The role of data/code archives in the future of economic research. *Journal of Economic Methodology 15*(1).

Banks, D. (2011). Reproducible research: A range of response. *Statistics, Politics, and Policy 2*(1).

Bourne, P.E. (2011). A few RDAP thoughts based on experience with the RSCB Protein Database. Paper presented at the Research Data Access and Preservation (RDAP) meeting, Denver, Colorado. Retrieved from http://www.slideshare.net/asist_org/bourne-rdap

Coles, S., Carr, L., & Frey, J. (2007). *R4L: The repository for the laboratory.* R4L Final Report, v0.1. Retrieved from http://r4l.eprints.org/publications/docs/R4Lfinalreport.pdf

Franklin, T. (2011). *What makes a good repository?* Comment on JISC-REPOSITORIES Archives email list. Retrieved from https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=JISC-REPOSITORIES;9975d539.1103

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday 16*(2). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3316/2764

Jasny, B.R., Chin, G., Chong, L., & Vignieri, S. (2011). Data replication & reproducibility. *Science 334*(6060). Retrieved from http://www.sciencemag.org/content/334/6060/1225.full.pdf

McCullough, B.D. (2007). Got replicability? The Journal of Money, Credit and Banking Archive. *Econ Journal Watch 4*(3).

McCullough, B.D., McKitrick, R. (2009). Check the *numbers: The case for due diligence in policy formation.* Report by the Fraser Institute. Retrieved from http://www.fraserinstitute.org/research-news/display.aspx?id=12933

OECD. (2004). *Declaration on access to research data from public funding.* Retrieved from http://www.oecd.org/document/2/0,3343,en_2649_201185_38500791_1_1_1_1,00.html

Open Society Institute. (2002). *Budapest open access initiative.* Retrieved from http://www.soros.org/openaccess

Palmer, C.L., Cragin, M.H., Heidorn, P.B., & Smith, L.C. (2007). Data curation for the long tail of science: The case of Environmental Sciences. *Science 331*(11).

Panagopoulos, C. (2011). Social pressure, surveillance and community size: Evidence from field experiments on voter turnout. *Electoral Studies 30*(2). doi:10.1016/j.electstud.2010.10.005

Peng, R.D. (2009). Reproducible research and Biostatistics (with discussion). *Biostatistics 10*(3).

Peng, R.D. (2011). Reproducible research in Computational Science. *Science 334*(6060). Retrieved from http://www.sciencemag.org/content/334/6060/1226

Steinhart, G., Dietrich, D., & Green, A. (2009). Establishing trust in a chain of preservation: The TRAC checklist applied to a data staging repository (DataStaR). *D-Lib Magazine 15*(9/10). Retrieved from http://www.dlib.org/dlib/september09/steinhart/09steinhart.html

Stodden, V. (2009). Enabling reproducible research: Licensing for scientific innovation. *International Journal of Communications Law and Policy 13*. Retrieved from http://www.ijclp.net/files/ijclp_web-doc_1-13-2009.pdf

Stodden, V. (2011). *Trust your science? Open your data and code!* Amstat News. Retrieved from http://magazine.amstat.org/blog/2011/07/01/trust-your-science/