Julia Kim
Data-PASS (10/25/2010)

Digitized data was initially created in the guise of tabulation machine results for the 1890 United States Census, however since the 1930s, less than half of social science research data has been meaningfully archived (Gutman et al. 319). Through a combination of lack of incentives (stakeholder interests), administrative burdens, and technological inadequacies; social science data has been, literally and figuratively, shredded and trashed. Data-PASS, or the Data Preservation Alliance for the Social Sciences, was initially created in 2004 to address and, especially now, to prevent these kinds of failures.

The Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIP), supported by legislation and funding in 2000, launched over 6o different collaborative projects to preserve at-risk digital information (Cruise and Sandore 301-302). Data-PASS emerged in 2004, encompassing the needs of the Social Science community, a field that has been historically low in its culture of sharing and publishing of its data (Gutman, "Preserving") Data-PASS is a federated consortium, led by six foundational archives: four university-based archives included ICPSR (the lead organization) centered out of University of Michigan (ICPSR is itself a consortium of over 550 universities world-wide), the Murray Research Archive and the Virtual Data Center (both at Harvard University), the Roper Center for Public Opinion and Research (at the University of Connecticut), and the Odum Institute (University of North Carolina at Chapel Hill). Data-PASS also has the electronics record management division of NARA as one of its founding partners. They have been key in advising and advocating legislation mandating data sharing. While the archives differ in their size, collection coverage, and collection materials, the complete collection is accessible at each node through a single common catalogue (Gutman et al. 315-316, 322-325).

Data-PASS's aims were simple but very comprehensive and, given the legal and administrative complications encountered, quite daunting : 1) To identify at risk social science data and to 2) Negotiate for its acquisition. "Social Science data" was very broadly defined and, data could take the form of interviews with traditionally unrepresented women (Murray archive), presidential polling data from the South, and even surveys on Tonya Harding (Guttman, "From Preserving"). Initially, Data-PASS targeted a few major categories of data that were of

institutional interest to the repositories: 1) Surveys and administrative data collected by and for the U.S. government (NARA). 2) Public opinion polls conducted by well-established institutions (Roper and Odum) 3) Research data collections supported by NIH and NSF grants 4) Research data collected by non-government organization or private organizations (ex: RAND corporation) (Gutman et al. 321). All data considered for ingestion was already digitized, whether born digital and simply in need of refreshing or migrating to archival, non-proprietary formats (data in SAS, SPSS or STATA migrated to ASCII). Not all data was purely quantitative; the Murray archive, for example, has a rich holding of archived interviews. Decisions on appropriate material to ingest were discussed in bi-weekly telephone meetings held by the Operations Committee. Even though ICPSR is the lead organization, each institution, however small, was granted a voice and a minimum of 1 delegate in the Steering Committee (Altman 344). Every aspect of Data-PASS involved regular coordination and communication among repositories. Cultivating relationships with private and non-governmental organizations was also especially important in order to negotiate data acquisition (Gutman et al. 326).

Much more than a long-term storage space, Data-PASS, from its inception, aimed to make sureits data was accessible through its shared catalogue (accessible through each repository). Data-PASS follows best practices and standards; it is OAIS compliant in its transparency, explicitness, adoption of non-proprietary software formats (when possible). For example, Data-PASS allow you to look up data (by its DOI- or digital object identifier number) to find and download the XML metadata file associated with the data. Data-PASS's basic architecture follows the standard LOCKSS model developed at Stanford, but in 2008, D-PASS created a PLN (private LOCKSS network) for its members that forgo HTTP or FTP use in favor of OAI-PMH. With LOCKS as its digital preservation internet "appliance," 6 - 7 copies of data are kept at various member repositories (at a granular level, which eases automated correction and data recovery), but nodes (and repository sizes and commitments) are asymmetric. In keeping with LOCKS strategies, there is no centralized single authority; it is peer to peer. It also uses CLOCKSS mechanisms to crawl and cache sites as another replication strategy. Both of these internet bases appliances require relatively little technical oversight to run, in keeping with D-PASS's desire to keep entry barriers low for institutions. Also in keeping with that, D-PASS attempts to keep submission information requirements (In its SIP) minimal to lower administrative burdens. It requires a minimum of 12 meta-data fields (modeled after Dub-Core).

Each member can be powered by the Dataverse Network (DVNm created in 2007), self-contained data archives that are all connected by the shared catalogue. (DVN, as of 2009, hosts over 140 virtual archives) (Altman 341). This freeware was developed and hosted by Harvard's Virtual Data Center. This Web application software imposed a minimal burden on Data-PASS partners. Interfaces for each of the members look very different from one another and members retain full dissemination controls over data. Query structure across the shared catalogue, however, remains the same. Essentially, search terms follow Boolean logic (powered by Lucene). There are some differences in query options across repository sites: while the Data-PASS site is limited to a single simple search field, ICPSR allows you to browse data collections through a "tree" structure outline, through a number of different search fields (Date, Distributor, Global i.d.), as well as Boolean searches.

In the case of the Murray archive at Harvard, Data-PASS catalyzed its transformation into a smoother running, more widely accessed, and much larger "living archive." Acquisitions increased four-fold ((Altman 339) after Data-PASS. Through the course of refining its practices, Data-PASS has continued to push for acquiring at-risk social science data at specialty archives like the Murray (there has been talk of acquiring Princeton's Culture Policy and Art National Data Archive (Gutman, "From Preserving"). Data-PASS, however, has also pushed for larger systemic changes in the basic data acquisition model. In its troublesome attempts to locate and acquire data from corporations, universities, and private researchers; confidentiality and ownership legal questions kept cropping up (Gutman et al. 328-329). As Gutman states regarding acquisition and use struggles, "...a lot of people can say no; almost no one has the authority to say yes "("Preserving"). Despite best efforts and practices, these problems have pushed Data-PASS to strongly advocate the creation of data-preservation and collections strategies with researchers and institutes before data collection has even begun. That is, they want data preservation to be a part of the plan from the very start, when studies are initially being designed. Data-PASS has also become a strong advocate for stronger legal mandates for researchers to systematically share their data. As of now, the weak legal binds for large NIH and NSA grant winners is just the beginning.

In conclusion, while Data-PASS has run out of its original funding and must work from grants, it has staked large-claim to truly be an inter-operable, “living” preservation archive that is accessible to many individuals and institutions at little to no cost.

Work-Cited

1. Altman, Micah. "Transformative Effects of NDIIPP, the Case of the Henry A. Murray Archive." Library Trends 57.3(2009):338-351. print.
2. Cruise, Patricia and Beth Sandore. "Introduction: The Library of Congress National Digital Information Infrastructure and Preservation Program." Library Trends 57.3(2009):301-314. print.
3. Guttman, Myron P. Preserving At-Risk Digital Social Science Data: The Data-PASS Project . Library of Congress. IN. 26 January 2007.
http://www.loc.gov/today/cyberlc/feature_wdesc.php?rec=4018
4.Guttman, Myron P., et al. "From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data." Library Trends 57.3 (2009).
5. Data-PASS: Data Preservation Alliance for the Social Sciences. University of Michigan, Fri. 22. Oct.2010.
6. "Data-PASS Metadata Requirements." 2007. pdf.
http://www.digitalpreservation.gov/partners/datapass/high/data-pass_metadata_requirements2007.pdf
7. ICPSR: Inter-University Consrotium for Political and Social Research. Version 6. University of Michigan, Fri. 22. Oct. 2010.