FIPNet Comments

Martin Halbert, 6/25/15, San Francisco, at the Federal Information Preservation Network Meeting

I was honored to represent the perspectives of the University of North Texas to provide some introductory comments regarding the case for the FIPNet. The information of the United States government will always be of enduring importance and significance culturally, economically, historically, and scientifically because of the sheer scale of the government, its unique centrality to our nation, its simultaneously singular and plural institutional character, and our enormous and ongoing shared social investment in it as a country and a people. The UNT Libraries have a long-standing deep commitment to preserving and providing research level access to government information, and we invest heavily in this program for a number of reasons. I think the case for FIPNet begins with three broad issues that are simultaneously challenges and opportunities.

Challenge #1: The Funding/Priority Transition in Libraries from Print to Digital Information

UNT became a congressionally designated selective depository in 1948 (when all government information was printed), and was an active (but perhaps typical) member of the program until the advent of the Internet. In 1994, UNT became an electronic Gateway for the GPO; the Gateway Project (as many of you will recall) was developed during the early days of GPO Access to maximize free public availability of GPO resources through federal depository library portals. Although GPO ended its formal support for the Gateway Project in the year 2000, we at UNT had by that time become convinced that ensuring digital preservation of and access to government information was a critically important programmatic function for research libraries, and we began to explore and experiment with new related activities. In 1997, UNT became the second depository library in the nation to join the Content Partnerships Program of the FDLP, a program committed to ensuring permanent public access to electronic federal information. Most seminally in that year, the UNT Libraries, as part of the FDLP and in collaboration with the GPO, established the first ongoing academic program to provide permanent public access to Web sites and publications of defunct and vanished U.S. government agencies and commissions. This program was led by my associate dean Cathy Hartman together with others at UNT, and this collection became known as the "CyberCemetery" by users of the site. In 2006, the National Archives and Records Administration joined this partnership, in the process designating UNT as one of nine existing NARA Affiliated Archives, and the only non-military affiliate. The UNT Libraries were chosen to serve as the official host of the permanent online collection of the defunct Advisory Commission on Intergovernmental Relations and other agencies. As we observed the growing usage of these web archives, we quickly came to understand the need for ongoing funding for this program, the scholarly value of preserving federal agency information, and that the ephemeral nature of the Web certainly pertained to many other critically important areas of the government web-presence as well. We began to broaden our efforts consequently, notably by joining forces with a handful of other institutions to undertake what has become known as the End-of-Term (EOT) Crawl project.

Challenge #2: Building and Maintaining Digital Expertise and Infrastructure

Before undertaking the EOT Crawl project, our library had, over a decade, by means of federal, private, and university funding, built up a collection comprised of more than 12 million discrete digital objects. Roughly half of those files were created and maintained by UNT as the statewide Portal to Texas History. The Portal includes six hundred thousand titles, many of which are historical newspaper series which were digitized during our participation in the National Digital Newspaper Program. The Portal
content is the basis for our role as the DPLA service hub for the state of Texas, and many other nationwide efforts. In the course of those years of digital library development work, we were able to build up our technical expertise and infrastructure, without which we would have been unable to take up the EOT Crawl work which is far greater in scope and potential application than even the relatively large Portal to Texas History. The notion of a research library has long included the expectation of responsibility for a collection of millions of physical volumes; in the 21st century this notion has to evolve to include millions of digital objects. Most of the other organizations in this room have had some parallel track by which you built up expertise and infrastructure at scale, but I think we have to figure out ways of assisting other organizations through FIPNet by broadening the base of digital library expertise in the country and jumpstarting their progress more quickly and affordably.

**Challenge #3: Organizing Collective Action at Scale**

Most of you will be at least somewhat familiar with the EOT Crawl project, which has as its general aim the goal of periodically capturing and preserving the entire federal web presence every four years when the presidential election occurs and there may be wholesale administrative information changes. This project, undertaken since 2008 as a collaborative effort between the Internet Archive, the Library of Congress, the California Digital Library, UNT, GPO, and since 2012 Harvard, is a massive undertaking that requires coordination on a grand scale to crawl, at least minimally organize, and then preserve the thousands of web sites and millions of web pages, PDFs, and other content that comprise the .gov, .mil, and broad relevant swaths of .org and other domains that contain federal information. The first such collaborative crawl in 2008 produced a collection of more than 16TB and 160M URIs, the 2012 crawl was 31 TB and far more URIs. We have no doubt that the 2016 crawl will be far larger still. In both the 2008 and 2012 cases UNT was responsible for conducting focused agency crawls and deep web capture of specific databases. UNT exchanged all crawled data with partners, with UNT repositing a full set of all crawls. This complete set only resides at UNT and Internet Archive. We have since worked with the Internet Archive to extract and begin to analyze all historical .gov content within the Global Wayback Machine, a new 110 TB archive that we are now trying to wrap our heads around.

It is this question of how to provide access to such massive aggregations of content that now preoccupies us, and we think will form for years to come a centrally important collaborative research agenda for archives like ours of digital government information. No human being can read a hundred terabyte archive; researchers are now in desperate need of more effective tools for text mining/analysis, and shared methodologies for identifying patterns and making scholarly claims using massive digital archives. This need is increasingly understood by most researchers seeking to make sense of the historical Web, but we at the UNT Libraries and UNT iSchool believe that the scale, public ownership, and central importance of federal information call out this need in high relief. But preservation logically precedes access, someone has to aggregate information before researchers can access it.

Both tasks of preservation and access present enormous challenges to libraries and other cultural memory organizations, and clearly no one organization can hope to make much headway alone. Our hope in joining FIPNet is to participate in the creation of a like-minded community of institutions dedicated to addressing these challenges jointly, by pooling expertise, infrastructures, resources, and moral commitment. That is the grand challenge and opportunity presented by FIPNet, and we are very pleased that the day has come for this inaugural conversation. Thank you.