Against the Grain / February 2009

Libraries Need a Little Help From Their Friends

Libraries have traditionally taken an active role in collecting content to meet the needs of their local user base. This was a straightforward process in the print world, with vendors galore and, in the case of U.S. government documents, the FDLP. In the digital world, things are much murkier, the process a little more convoluted. The responsibility to collect and preserve content remains but the process is more challenging; on the open Web, there are no vendors to pull together disparate publishing streams or depository systems for easy inclusion into local library collections. On the Internet, libraries need to implement a more aggressive approach toward collecting Web-based materials as well as identifying new partners in their efforts — libraries must rely on the kindness of strangers and library fellow travelers.

One such fellow traveler to the government documents community is Carl Malamud. Malamud is an Internet and open government activist who runs the Website, public.resource.org. Since the U.S. government has been producing digital public domain government information, Malamud has been successfully shaking it free from government control and onerous access fee structures and making it more accessible to citizens. Malamud’s overarching goal is to release government information into the open so that others can build more advanced interfaces and facilitate better access to the workings of our governments.

His first campaign led to the creation of the Securities and Exchange Commission’s Edgar database of SEC filings and corporate disclosure documents (which has recently had a name change to IDEA). He has since, in his efforts to “open source America’s operating system,” set his sights on Federal and State Courts and case law, State and municipal codes, U.S. Copyright Office, National Technical Information Service (NTIS) videos, Government Accountability Office (GAO) legislative histories, and, of most interest to Government Libraries, the FDLP documents from the GPO.

The GPO is the official publisher of the U.S. Government and manages the FDLP. They publish and distribute to libraries publications from 21 federal agencies as well as such integral publications as the Federal Register, Congressional Record, Congressional Reports, Bills, documents and Hearings, Public Laws, Papers of U.S. Presidents and much more. GPO Access is built on an older technology called WAIS with a very primitive user interface and limited search capabilities. For that reason, Malamud, with the assistance and cooperation of the GPO, harvested GPO Access documents from GPO servers in late 2007 and made them accessible/downloadable via BitTorrent, Rsync, HTTP and FTP. Those documents comprise 200+ gigabytes of data from 1991-2007 amounting to 5,177,003 PDF pages, 54,600 GAO Reports, 448,496 Congressional Reports and more. It’s these GPO documents upon which the USDocsPLN has so far focused.

Current Status

The USDocsPLN is now up and running. The 200+ gigabytes of digital documents have been downloaded from Malamud’s site (http://bulk.resource.org/gpo.gov) and distributed among the 15 partners in the project, where they will be preserved within the LOCKSS network. This was an extremely cost-effective project as 1 terabyte (which equals 1,000 gigabytes) of storage is now below $200, hardware is typically less than $1,000, and there is only minimal administrative cost once the LOCKSS box has been configured. The group will continue to evaluate and add to the network other