Public Access to Large Data Sets in a Depository Library

by Juri Stratford, Government Documents Department Shields Library University of California, Davis*

In the United States, depository libraries receive federal publications under the Depository Library Program as described under Title 44, Chapter 19 of the United States Code. Depository libraries act as custodians for federal publications in exchange for providing public access to those government publications. While the Census Bureau started experimenting with the distribution of data on CD ROM in 1985, the full scale depository distribution of CD ROMs started about 1990.

There is no single agency within the Federal government responsible for coordinating the format of the data distributed. While the Government Printing Office is the agency responsible for distributing the depository CD ROMs, GPO doesn't fulfill the role of a publisher. The format of the data files and the software to access those data files, if any, is the decision of the data producer, as is the decision whether or not to provide a given data product to depository libraries.

The early momentum behind the depository distribution centered around MS DOS. The Census Bureau began using the dBase format for their files beginning with Test Disc 2 distributed to depository libraries as an experiment in 1987. Most Census Bureau CD ROMs are still distributed in dBase format. It was easy enough for depository libraries to provide access to the CD ROMs either using programs produced by the Census Bureau when available, or in other instances using dBase.

Given the nature of the depository library program, depository libraries are reactive rather than proactive. Depository libraries receive publications based upon item selection surveys. Each item number represents a class of documents or datafiles. Depository libraries frequently do not know the file format of, and software access to, the CD ROM products before they are selected. This means that as their datafile collection takes shape, they must develop access strategies after the fact.

At the University of California, Davis, we have made specific decisions regarding the types and levels of public service that we can provide for datafiles. These public service activities include loaning the CD ROMs; making the CD ROMs available on public microcomputer stations; making the CD ROMs available via the network; and providing basic extraction of data subsets for end users. We have limited our service to the provision of files, either as created by the data producer, or custom subsets. While we use a variety of applications to produce these subsets, we do not provide any access to analytical or statistical software, including mapping software. So, for example, while we provide mapping data on a regular basis, we do not produce maps.

The original depository CD ROMs distributions did not include front end software. Our options at that point were to work with the CD ROMs ourselves using dBase or to loan the CD ROMs. Loaning the CDs was not seriously considered at this time because very few people had access to CD ROM drives. Our solution was to work with the datafiles, create subsets, and distribute the data on floppy diskettes.

As the data producers began to develop some user-friendly front ends to their data, we began to make these CD ROMs available on public microcomputers. At present, we have five MS DOS microcomputers that can be used by the public to access depository CD ROMs. Four of these CD ROM stations are in a public area. We provide access to about sixty CD ROMs on these four public stations. The primary consideration for mounting a particular CD on these public stations is our subjective evaluation of the front-end software provided by the data producer. We mount only files with appropriate front-end software for unmediated public use on these machines.

The fifth CD ROM station is a computer in the back of the department that can only be accessed when the department is open. Researchers can use this computer to work with CD's that are not available on the four PC's in the department's public reference area, for example, lesser used CD ROMs such as the Census CD ROMs for other states or incorporating complex software such as the National Health Interview Survey CD ROMs. This computer is also the only computer with dBase software. Where extractions are too complex to be done easily or efficiently with the vendor produced front end, we provide extraction services using dBase software at this station.

We did not have a formal policy to loan CD ROMs until last year when we instituted a three day loan period for CD ROMs that were not accessed on the public CD ROM stations. We are usually able to loan most of our CDs due to our arrangement with the Law Library . As a second depository on the same campus, the Law Library selected the CD ROMs and transferred them to our department. Two other developments influenced this decision. One was the increase in the number of CD ROMs that we have received in formats not specific to MS DOS environments. These include flat files, microdata files, and geographic or image files. As we don’t have appropriate software in the department to work with these types of data, we allow users to take them to other sites that do. Second was the increase in PCs equipped with CD readers. While it was once rare for end users to have their own CD reader, it is now quite commonplace, and many users would prefer to work with the data on their own systems.

The geographic datafiles represented the largest category of files‘ for which we did not offer any computer-based access within the department. We have a large number of departments and labs on campus working with geographic files; and while we don’t have the facilities to provide GIS services in our department, we are the largest archive of raw geographic data on campus. This includes the Census Bureau’s 1990 and 1992 TIGER Line Files, and the U.S. Geological Survey’s Digital Line Graph series at 1:2,000,000 and 1:100,000. We are also anticipating the receipt of approximately three hundred fifty CD ROMs representing the U.S. Geological Survey’s Digital Orthophotoquads for California, geographic image files, in JPEG format.

Our network approach was not part of some grand plan, but rather a large number of circumstances coming together at all at once. We decided in Spring 1994 to ask for a grant for more computing equipment to support access to geographic data files. The Census Bureau was starting to distribute the Landview software with the TIGER Line Files, and we felt that we could not work with this software on an 80386, our most powerful microcomputer. As 80486s were becoming common, we thought that we would ask for this. Our administration increased our request for equipment; it was not coming out of their budget. We ended up with a Pentium, a larger monitor, and a color printer. However, when the equipment arrived, it was still not obvious what software we would use or even what public access we would provide to the equipment.

Before the equipment arrived, the staff support person that I had for computing left to go to another department on campus. We decided that, rather than hire new staff, we would hire a student. When we were able to hire a student who was knowledgeable about both UNIX and networking, we developed our strategy around this person. So we tried Linux on our new equipment. Linux is a freely available UNIX system for Intel based PC’s. We decided upon Linux for several reasons. First, because of the complete network support offered; and second, because we had a student capable of installing and maintaining the system. We were also intrigued by the possibility of running GRASS on the system. GRASS is a free GIS system developed by the U.S. Army. GRASS is available for several platforms, including Linux, and is supported by other GIS projects on campus. While we have not worked with GRASS at this point, we are cooperating with GIS labs on campus that use GRASS, and are now examining GRASS as an extraction tool for geographic data.

Our first objective was to provide anonymous FTP access to the system. We started out with two triple speed CD ROM drives, and later replaced them with four double speed CD drives as we decided that the slower CD drives were adequate for throughput on the network. Our most heavily used CD’s that could best be accessed in this manner were our 1992 TIGER Line Files for California; these are distributed on three CD’s. With four CD drives, we have dedicated three CD ROM drives to TIGER leaving one drive free for other data. Since implementing anonymous FTP access to the TIGER Line Files, we have also made the files available via the World Wide Web using an http front end to the anonymous FTP access. Our CD ROM drives are also available to local users via NFS on an experimental and still restricted basis.

We focused on network access to our large depository data sets for several reasons. First, we were not able to provide adequate access to the data within the department, and several appropriate computing facilities were available on campus. Second, there were competing demands on campus for the TIGER Line Files that could not entirely be met by loaning the CD ROMs: several different researchers would request the same files at the same time, and the researchers frequently did not have adequate access to CD ROM drives in the GIS facilities. Third, network access was readily available in the building. Fourth, appropriate software, i.e. the Linux operating system, was freely available; and finally we were able to hire an experienced student to implement the system.

So far, we have been able to distribute the raw TIGER files to campus users via anonymous FTP. We have been able to make additional CD ROMs available on the system as necessary. We have been able to upload large data extractions created on the 80386 system, the fifth public access microcomputer described above, onto the UNIX system for anonymous FTP access. And finally, we have started to develop an http interface to our anonymous FTP system for access via the WWW.

The networked access to the depository CD ROMs is still very much an experiment, but we have been satisfied with the success of the project so far. We still need to implement a formal system of communication via email to rotate CD ROMs through the fourth CD ROM drive, and once that we conclude that the system is stable we need to increase our efforts to publicize the system.

We do not intend to make a large number of data files available on the network permanently. Our objective in developing this system has been to use the network to serve a local community of data users. However, we have no objections to outside use to the extent that outside use of the system does not compete with local user needs.

* Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.