SAND REPORT

SAND 2002-0107
Unlimited Release
Printed January 2002

Long-Term Spatial Data Preservation and Archiving: What are the Issues?

Denise R. Bleakly

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

This country has spent billions of dollars in the collection of geospatial data, and we assume that "someone else" is taking care of the data. In doing this research, several alarming examples of important digital data being lost over time were found.

The most complex and well documented was the Canada Land Data System (Brown and Comeau, 1999). The Canada Land Data System (CLDS) was the first GIS system in North America, and it was designed to map information related to the Canada Land Information (CLI) System. This system began development in 1963 and collected information to develop a nationwide land database as the basis for multidisciplinary land-use planning. The CLI consisted of multiple coverage, land-use capability maps for agriculture, forestry, recreation, wildlife, and land-use activity. The CLI was the largest single land capability assessment done in any country.

This system was developed as a proprietary IBM mainframe product with the PL/1 computer language. By the mid 1970s, approximately 3500 maps were available from the system. In 1994, the CLDS was discontinued because a program review led to a reorientation of the department responsible for the system. The CLDS was archived.

In 1979, the National Archives of Canada, reviewed the 1974 decision to archive the CLDS and determined that the CLDS was still of archival value, but the software was obsolete and would have required considerable effort and resources to make it operational. The collection consisted of 2965 nine-track tapes.

The decision in the mid-1990s to try and restore the data but not the software created many technical problems. Many of the tapes had suffered "stiction' problems and were literally falling apart. Stiction happens when the magnetic tape sticks to the read-write head of the tape drive. Stiction did not occur evenly across the tape. Eventually, some data layers could be copied from tape to disk, but during verification procedures, there were many files filled with data errors. It seemed that the restoration of the CLDS was unfeasible.

It was then decided to try and extract the data and convert them to more modern GIS software platforms. New data translators (essentially system emulators) were written to take the data from the original data structure to convert them to Arc/Info Generate files, Digital Line Graph files, and SPANS vector files. This was not an easy task. It was not just a matter of simply reformatting the data from one format to the other. Specialists were required to understand the underlying data structure of the data and to write the data to a new data structure without the loss of the data integrity. The data conversion process has evolved over time to a 16-step process. Several data sets were unrecoverable – some waterfowl data and land use data in the Eastern Townships.

Much of the data was ultimately recovered, but much work remained to make the data useful to current geospatial data users. The first sets of recovered data were released to the public in 1998. Since the data were converted to ArcInfo format, the response to these data layers has been "overwhelming." These maps are currently available for free

on the Canadian Geospatial Data Infrastructure GeoGratis web site: http://geogratis.cgdi.gc.ca/frames.html.

Although this data-recovery effort has been a success, it was very costly, very time consuming, and very technically challenging, and there was indeed some loss of data.

Another example of lost digital data came from the 1960s decennial census. The Commission on Preservation and Access (1996) briefly described the loss of this valuable data set. In 1976, during a review, the National Archive identified seven series of aggregated data from the 1960 census files as having long-term historical value. The data resided on UNIVAC Type II-A tapes. By the 1970s this tape drive was long obsolete. The Census Bureau had significant technological problems preserving the datasets. It was able to successfully migrate the data to new industry standard tapes, but they did not analyze the integrity of 1.5 million records of aggregated data. Instead, the Bureau chose to label the tapes with a warning that only two systems existed in the world that could read the data. These two systems were eventually eliminated because they were not labeled as the only equipment available to read these tapes. Consequently, the ability to read these tapes was completely lost.

What Makes Geospatial Data Unique?

Geospatial data are a special subset of digital data. Geospatial data represent many facets of phenomena on the earth and are stored as points, lines, polygons, regions, volumes, and grids. A point may represent a well or sampling location, a line may represent a boundary, a fence line, or a road; a polygon could represent a containment cell or area that should be left undisturbed. Regions, volumes, and grids are methods that could be used to represent areas of subsurface contamination or areas of groundwater concern. Geospatial data stored in a GIS usually have relationships between objects stored as part of the data structure. The power of geospatial data is the ability to derive new data from relationships between data layers (Zaslavsky, 2001).

Geospatial data are multi-scaled and have multi-resolutions. In any particular geospatial data set, the data may be valid for a range of scales. For example, a data set representing a detailed engineering diagram for a containment cell may have a valid scale of 1 inch on the map equals 50 feet on the ground, or 1”=50’, while a map of a state showing all the locations of groundwater concerns may be at a scale of 1 inch on the map equals 50 miles on the ground or 1”= 50 miles. This factor of digital geospatial data makes it very difficult to preserve because the range of valid scales and resolutions for a particular data set are key to knowing the appropriate use of a particular data set in the future.

Geospatial data can be both current and historical, and the large amounts of geospatial data that could be preserved and archived could prove to be very valuable to future researchers looking for long-term changes in the environment or ecosystems. In the case

References

Brand, Steward, 2000. Written on the Wind. *Civilization.* October/November 1998, pp. 70-72.

Brown, David, and Comeau, Mike. 1999. Restoration of the Canada Land Data System. *Association of Canadian Map Libraries and Archives Bulletin*, Number 106, Autum 1999, pp. 42-52.

Commission on Preservation and Access and Research Libraries Group, Inc., (Commission), 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* A copy of this document can be found at http://www.rlg.org/ArchTF/tfadi.index.htm.

Federal Geographic Data Committee (FGDC), 2001. *Managing Historical Geospatial Data Records: A guide for Federal Agencies.* Internet Fact Sheet: http://www.fgdc.gov/nara/hdwgfsht.html.

ICF Kaiser Consulting Group (ICF). 1998. *Managing Data for Long-term Stewardship*. Report prepared for EM Office of Strategic Planning and Analysis, March 1998. A copy of this document can be found at: http://lts.apps.em.doe.gov/center/reports/doc1.html

O'Connor, Maura, 1996. *Digital Spatial Data: Accessible, But for How Long?* Addresss given a the National Library of Australia at the Mapping Sciences Institute of Australia 1996 Biennial Conference. A copy of this address can be found at http://www.nla.gov.au/nla/staffpaper/oconnor1.html.

Piwowar, Joseph M., 1998. Putting Your data Out to Pasture. *Cartouche*, No. 29 Spring 1998. A copy of this document can be found at http://www.watleo.uwaterloo.ca/~piwowar/Think/Archiving.html.

Stewart, Elenor, and Banks, Paul N., 2000. *Preservation: Issues and Planning*, Chapter 18, Preservation of Information in Non-paper formats. American Library Association, Chicago and London. pp323-342.

U.S. Department of Energy (DOE), 1998. *Accelerating Cleanup: Paths to Closure*, DOE/EM-0362, June 1998. A copy of this document can be found at http://www.em.doe.gov/closure/final/index.html.

U.S. Department of Energy (DOE) 2001. Report to Congress on Long-Term Stewardship Release No. R-01-025 Release Date: January 19, 2001. A copy of this document can be found at: http://lts.apps.em.doe.gov/center/reports/jan01rtc.html.

Zaslavsky, Ilya, 2001. Archiving Spatial Data: Research Issues. San Diego Supercomputer Center Technical Report TR-2001-6, January 18, 2001.
A copy of this report can be found at http://www.sdsc.edu/TR/TR-2001-06.doc.pdf.