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.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2002-0107
Unlimited Release
Printed January 2002

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

Denise R. Bleakly
Site Closures
Sandia National Laboratories
P. O. Box 5800
Albuquerque, NM 87185-0763

Abstract

The Department of Energy (DOE) is moving towards Long-Term Stewardship (LTS) of many environmental restoration sites that cannot be released for unrestricted use. One aspect of information management for LTS is geospatial data archiving. This report discusses the challenges facing the DOE LTS program concerning the data management and archiving of geospatial data. It discusses challenges in using electronic media for archiving, overcoming technological obsolescence, data refreshing, data migration, and emulation. It gives an overview of existing guidance and policy and discusses what the United States Geological Service (USGS), National Oceanic and Atmospheric Administration (NOAA) and the Federal Emergency Management Agency (FEMA) are doing to archive the geospatial data that their agencies are responsible for. In the conclusion, this report provides issues for further discussion around long-term spatial data archiving.

Key words: Data Management, Information Management, Geospatial data, Archives, GIS, Long-Term Stewardship, DOE.

Contents

Nomenclature

CAD	Computer Aided Design
CLI	Canada Land Information (System)
CLDS	Canada Land Data System
DOE	Department of Energy
ER	Environmental Restoration
ESRI	Environmental Systems Research Institute
FEMA	Federal Emergency Management Agency
FGDC	Federal Geographic Data Committee
GIS	Geographic Information Systems
GPS	Global Positioning System
LTS	Long-Term Stewardship
NCDC	National Climatic Data Center
NOAA	National Oceanic and Atmospheric Administration
NSLRDA	National Satellite Land remote Sensing Data Archive
NSDI	National Spatial Data Infrastructure
SDTS	Spatial Data Transfer Standard
SVG	Scaleable Vector Graphics
USGS	United States Geological Survey
USGS/EDC	USGS EROS Data Center
VML	Vector Markup Language

Introduction

During the last twenty years, it has been estimated that the Department of Energy (DOE) has spent approximately $147 billion dollars performing environmental restoration activities at hundreds of sites across the United States (DOE: 1998). These sites were associated with the development and maintenance of the U.S. nuclear arsenal. Recently, the DOE has made the determination that for many of these sites, it is neither practical nor safe to remediate them to contamination levels that allow for unrestricted re-use of the land. The DOE's program of Long-Term Stewardship (LTS) has been set up in response to the US Congresses' call for understanding the long-term consequences of maintaining sites and monitoring sites that will be entering into the LTS program. The magnitude of the LTS program is large-it is expected to cost $101 million dollars annually by 2050 (DOE, 2001, pp.3-21).

One aspect that has not been considered as part of the LTS planning is the long- term cost of maintaining digital records of where these sites are, what contaminants were left at the sites and what levels of contaminants; the human health risk at these sites; and when or if the land could ever be re-used. The long-term preservation of digital records has been the topic of study by librarians and archivists for the last twenty years or so (Brand, 2000; The Commission, 1996), and currently, there are no easy answers as how best to guarantee that digital data will last into the future without active management to preserve the data.

One special subset of digital data of particular concern is digital geospatial data. Geospatial data has to do with the location of things on earth. Paper maps are a form of spatial data; other forms of geospatial data are electronic maps and drawings and data stored within a Geographic Information System (GIS). As part of the DOE's Environmental Restoration (ER) program, very large spatial data bases have been created both in computer aided design (CAD) formats and as GIS data. These large digital datasets have many of the data elements that will be needed for the future as DOE enters into LTS. Geospatial data are unique in the digital world because real-world phenomena are stored as points, lines and polygons; and relationships between these entities are stored as part of the electronic data structure. Trying to capture this data in paper format for archiving purposes falls short - the very nature of these relationships between data elements is made static in the development of paper maps. This presents a true challenge for DOE as many of these sites move into LTS.

This report discusses the issues surrounding the long-term archiving of digital geospatial data, present the challenges facing other branches of the federal government examining this complex issue, present findings from a survey of federal agencies and discusses possible paths forward for DOE sites moving into LTS.

Data Archiving

Records Management, Information Management, and Archivist personnel have been dealing with issues concerning the long-term preservation of digital data for the last 30 years, and in that time, there has been a shift from primarily paper records to electronic records (Stewart and Banks, 2000). This shift from paper records to digital records has caused serious problems in information management. Some of these issues for information preservation are changes in media, technological obsolescence, data refreshing, data migration, and data emulation. Each of these is discussed in detail below.

Issues for Digital Data Preservation

Media

While records were primarily paper in nature, it was relatively easy to estimate how much storage space was necessary and to determine the costs for preservation. However, now that more records are electronic, it is becoming increasingly difficult to estimate data preservation costs and what it will take to manage the records. One of the biggest concerns is with the media. Books and other paper media have an expected life span of fifty to one hundred years or more.

Most newer digital media are much less robust than printed books or paper documents because (Stewart and Banks, 2000) of the following:

- They are less chemically stable than even poor quality paper,
- They deteriorate more rapidly even when stored unused in good environments,
- Digital data are machine dependent that is they must move within machines to provide their information. Just reading the data incurs wear on the media,
- They are totally system dependent for retrieval of their information. When the system (either hardware or software or both) is no longer sustainable, the information will be lost unless it has been migrated to a newer system,
- Digital information technologies rely on ever greater data packing densities, making the information ever more vulnerable to large losses from small incidents,
- Failure of many newer digital media is often unpredictable and sudden, and may result in total loss of the information,
- For many newer media types, there is little experience with their maintenance and preservation.

Currently, there are three basic classes of archive media: magnetic disk (e.g., computer hard disks, removable disks); optical disks (CD-ROM and DVD-ROM); and tape (e.g., DAT and Exabyte) (Piwowar, 1998). Although there are variations on each one of these media types, these are currently the only viable options to be used for electronic data storage. One caveat is that, even under the best storage conditions, digital media can be fragile and have a limited shelf life (Commission, 1996).

Technological Obsolescence

In 1996, the Commission on Preservation and Access published a document, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* This commission focused on many of the challenges facing the United States in preserving its cultural and informational memory.

It reported that in breathtaking cycles of 2 to5 years, new devices, processes and software are replacing the products and methods used to record, store and retrieve digital information. Given that speed of change, the storage medium could outlive the availability of readers for that medium. The classic example is the 8-track tape.

There is a push to preserve the physical media that the data are stored upon. Many experts believe that "technological obsolescence represents a far greater threat to information in digital form than the inherent physical fragility of many digital media" (Commission, 1996).

There are other issues surrounding technological obsolescence that need to be accounted for in the digital data preservation discussion. As hardware and software systems evolve, there has generally been a lack of compatibility between hardware and software platforms. In the spatial data realm, this has become less of an issue as hardware and software vendors have been migrating to a few sets of de facto data sharing standards such as DXF, Shape, MiFF, TIFF, etc. However, there are proprietary data formats that are still prevalent.

Another large issue is that as software evolves, there generally is a lack of "backward compatibility." This is evident in 2001, with the Environmental Systems Research Institute's (ESRI) new ArcView 8.1 software. Earlier scripts and automation tools are no longer readable in the new version of the software.

In, *Preservation: Issues and Planning* (Stewart and Banks, 2000: p. 324), the authors note these less obvious concerns about technological obsolescence:

- Accessibility to digital information depends entirely on intricate edifices of hardware, operating systems, applications software, and storage media.
- Most such systems are heavily proprietary, which leaves those concerned with long-term preservation at the mercy of the market place.
- Changes in technology are almost wholly driven by business and consumer forces; libraries and archives have virtually no influence on these developments.
- Although there are many crucial standards, both formal and de facto, in the digital domain, developments in technology often move faster than the standards process.

Data Refreshing

One way many information specialists are handling the issue of technological obsolescence is simply to copy data from an older medium to a newer medium. This is termed "data refreshing." In theory, this is easy- you set up a schedule, and you just copy the data from one medium/format to another. However, this does not guarantee that the information with be usable with new software versions. Data refreshing can generally be carried out without loss of data because it is done on a bit by bit basis. However, there is always the risk of loosing data or functionality.

Data refreshing is a long-term solution for data preservation "only as long as the information is encoded in a format that is independent of the particular hardware and software needed to use it and as long as there exists some kind of software to manipulate the format in current use" (Commission, 1996).

Data Migration

Another way archivists are handling long-term electronic data archiving is by data migration. Data migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation.

Data migration tries to retain the ability to display, retrieve, manipulate, and use the information. Migration includes data refreshing, but is different, in that it tries to move the information forward by migrating forward not just the data but also operating systems and specific software codes. Data migration is time consuming, costly, and much more complex than simple data refreshing (Commission, 1996).

Emulation

Another approach that has been recently discussed (O'Connor, 1996) is emulation. Emulation, is a software tool that makes other software, act as if it were something else. It is now being tested for reading early digital text formats (remember Wang word processors?); however there are no long-term studies to determine if this approach has any long-term viability.

Data Archiving versus Information Preservation

There is a significant difference between data archiving and information preservation. Simplistically, data preservation is concerned with the long-term retrievably and use of information within its original data context.

Many data archiving programs focus on refreshing the data or copying it to newer media. Data copying can ensure that the information is archived on the newest media, but lose the integrity of the information. Digital data can be well archived but not well preserved.

Data archiving may facilitate preservation but not ensure it. Stewart Banks (2001) says the following:

> To the degree that there is a twenty or thirty-year time lag between the creation of records and the time they are accessioned into archives, and the accessibility of most desktop media is five to ten years, archives will only be able to hope to preserve digital information by taking an active role in its maintenance early in its life cycle.

Data Storage Versus Data Access

The Task Force on Archiving Digital Data defined digital archives strictly in functional terms as repositories of digital information that are collectively responsible for storing and ensuring through the exercise of various migration strategies, the long-term accessibility of digital data. Digital archives are distinct from digital libraries in the sense that digital libraries are repositories that collect and provide access to digital information, but may or may not provide for the long-term storage and access of that information (Commission, 1996).

There is an increasing discussion concerning the concept of data warehouses or data clearinghouses. In the geospatial data community, many of the practitioners are using "data warehouses" or data "clearinghouses " synonymously with data archiving, yet the way they are being used suggests a digital data library rather than a true archive.

Many federal, state, and local governments are creating these geospatial data warehouses as a way to access vast amounts of federally funded geospatial data.
One portal to the National Spatial Data Infrastructure (NSDI) data clearing house can be found at the National Geospatial Data Clearinghouse
http://130.11.52.184/. This is more like a collection of data sources, rather than an archive. It provides access to hundreds of sites that have geospatial data on-line.

However, the National Satellite Land Remote Sensing Data Archive (NSLRSDA) is an example of both a data clearing house and a true archive:
http://edc.usgs.gov/programs/NSLRSDA.html.

Long-term Costs

No one yet has been able to determine the total costs of preservation of digital data. Most of the research to date (Commission, 1996; Stewart and Banks, 2000) has determined that preservation of digital data will be expensive, primarily because data preservation is a manual process and is very labor intensive.

What makes digital data expensive to maintain? First and foremost, knowing what to migrate and preserve. Subject matter experts and information professionals will both need to be involved in the preservation process -- the subject matter experts for determining what the topic is and how relevant to LTS the particular data are, and information professionals to know how to preserve the data. On the surface, it seems that this should be rather straight forward; however, unless a particular data set has detailed metadata (data about data), there could be a point in the future that digital data sets might loose the context with which they were important.

Migration of the data is a very costly endeavor. The time spent copying data from one file type to another is tedious and time consuming. Further checks of the integrity of the data and checking for errors is an additional time-consuming but necessary task.

Another cost is building and maintaining indexes to archived information. As archives grow, the indexes to the information become critical to the management of these data. Again, building indexes to data, for the most part is a manual, time consuming process.

One option the National Oceanic and Atmospheric (NOAA) is taking is creating an archive of hardware and software (Davidson, 2001, Appendix B questionnaire). For the time being, they are keeping at least one of each type of data reader. In addition, they are keeping earlier versions of software to be able to read older media types. Most libraries and archives are not so fortunate. Stewart and Banks (2000) noted, most "general research libraries, archives and special collections departments may not be able to support the infrastructure-- specialized equipment, personnel and storage facilities -- necessary for long-term preservation of newer media".

Several authors (Stewart and Banks, 2000, Commission, 1996 and O'Connor, 1996) noted that one of the major concerns about long-term digital data archiving will be its long-term costs. The concern can be summarized by this quote from the Commission on Preservation and Access (1996):

> The cost and complexities of moving digital information forward into the future raise our greatest fear about the life of information in the digital future: namely, that owners or custodians who can no longer bear the expense and difficulty will deliberately or inadvertently, through a simple failure to act, destroy the objects without regard to future use.

Examples of Lost Digital Data

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 of LTS for DOE, these historical geospatial data may help in understanding long-term changes in land status/re-use or possibly long-term contaminant migration problems.

In addition, geospatial data can be in multiple formats: in the form of aerial photos, maps, surveys, global positioning system (GPS) data, CAD files, GIS data, etc. This poses a dual problem for long-term archiving, in that both developing and maintaining archives of digital data and paper media must be considered.

Why Are Geospatial Data Difficult to Archive?

Geospatial data pose special problems for preservation and archiving. "Spatial collections require special interfaces for querying and representation. Typically, spatial data are stored in proprietary system-specific formats and visualized with GIS, CAD or remote sensing software" (Zaslavsky, 2001).

There are many proprietary software platforms. Examples are ESRI Arc/Info and ARCGIS, Intergraph, AutoCad, MapInfo, and ERDAS. Although many of these systems are moving towards some level of interoperability, data collected in these proprietary formats are often difficult to read without translation software.

Another issue is that spatial data are stored in many different formats – Digital Raster Graphics (DRGs), Digital Orthophoto Quadrangles (DOQQs), Digital Elevation Models (DEMs), and Digital Line Graphs (DLGs), as well as proprietary file types. Currently, there are no uniformly accepted industry-wide standards for data formats or data exchange formats. An early attempt to develop a uniform standard was the Spatial Data Transfer Standard (SDTS), but this format has been difficult to implement because of industry's reluctance to create easily usable translators. The U.S. Geological Service (USGS) is the main government agency providing some data in this format. Information on the SDTS can be found at http://mcmcweb.er.usgs.gov/sdts/.

The Federal Geographic Data Committee (FGDC) (http://www.fgdc.gov/) is working on spatial data metadata standards and has recently joined the Open GIS Consortium (http://www.opengis.org/) to begin looking at broader guidance for geospatial interoperability standards. In addition, new developments in using platform-independent mechanisms for converting spatial data preservation formats into open-format Web presentation have recently been explored (Zaslavsky, 2001). Some of these include, Vector Markup Language (VML), Scaleable Vector Graphics (SVG), and a new G-XML project. The aim of the G-XML project is to create a protocol for encoding spatial data through extensions built upon XML. The goal is that G-XML will provide a method for freely accessing and using geographic information over the Internet. G-XML is a project of researchers in Japan and the U.S. The G-XML project web site gives a history of the project and current work (http://gisclh.dpc.or.jp/gxml/contents-e/). Currently, G-XML is in a testing phase, and the prospects for using G-XML as an archiving tool have not been explored.

Another critical issue for consideration is the volume of digital geospatial data that are generated and must be managed. Although physical storage media for these volumes of data are getting less expensive, large amounts of spatial data are expensive to maintain because of the indexing, management, and retrieval costs. For example, NOAA's National Climatic Data Center (NCDC) has over 1.1 **petabytes** of data that it currently maintains in its weather data archive (Davidson, 2001 personal communication). For comparison, a petabyte is 2 to the 50th power (1,125,899,906,842,624) bytes. A petabyte is equal to 1,024 terabytes. The USGS EROS data center National Satellite Land Remote Sensing Data Archive (NSLRSDA) is expected to have some **2,400,000 gigabytes** of data by the year 2005. To put this in perspective, a gigabyte is 2 to the 30th power (1,073,741,824) bytes. One gigabyte is equal to 1,024 megabytes. In any case, this is an ocean of information stored as bits and bytes on computer media!

Within the DOE complex, it is not unusual for geospatial data systems to contain gigabytes of information.

Existing Guidance

The FDGC has created the fact sheet "Managing Historical Geospatial Data Records: A Guide for Federal Agencies" (http://www.fgdc.gov/nara/hdwgfsht.html), which provides a general overview of federal agencies' responsibilities for properly creating data, documenting data with appropriate metadata, making data available through a node on the National Spatial Data Infrastructure (NSDI) and arranging for the appropriate disposition of the data.

This fact sheet discusses the current laws and has these major headings: What is the Law?, What Records Are Appropriate for Preservation?, How Do Agencies Document Information About Their Data Sets?; Sources of Information Should the Geospatial Data Set Be Saved? and Geospatial Data Base System Considerations. It is a good starting point for locating information and offices within the federal government for records information.

This fact sheet, under the heading of "How Do Agencies Document Historical Information About Their Data Sets?" goes into a brief discussion on the FGDC content standards for digital geospatial metadata.

Within the context of DOE, there are records retention schedules for most data types that DOE collects, but these schedules vary from project to project and subject to subject. Within the context of DOE's LTS program, there has yet to be defined any guidance concerning the archiving of geospatial data.

Geospatial Metadata

Metadata or "data about data" describe the content, quality, condition, and other characteristics of data. The FGDC approved the Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998) in June 1998 in an effort to standardize the characteristics of data so that data users can determine the data's fitness for their purpose.

According to the NSDI fact sheet on Geospatial Metadata (http://www.fgdc.gov/metadata/metadata.html), the major uses of metadata are:

- To help organize an maintain an organization's internal investment in spatial data,
- To provide information about an organization's data holdings to data catalogs, clearinghouses, and brokerages, and
- To provide information to process and interpret data received through a transfer from an external source.

Through the FGDC, users can download extensive information on the content standard for digital geospatial metadata, get training, and find software to compile metadata.

The ICF Kaiser report on long-term data management (ICF Kaiser: 1998) had a very detailed discussion on the various metadata format standards that the federal government is using for electronic media. It had a very good summary of the FGDC content standard for metadata:

> The FGDC metadata standards were developed pursuant to Executive Order 129063, which mandated that all federal agencies adopt these standards for all geospatial data sets created after January 1, 1995, in order to facilitate sharing of geospatial data sets among Federal agencies and with the public.
>
> These metadata standards specify the information content of metadata for geospatial data sets like maps, Geographic Information System (GIS) and Computer Aided Drafting (CAD) data sets, and other data files that contain information about where things are located. Metadata meeting these specifications are now required by OMB for any geospatial data disseminated to the public. While not required for non-spatial data, these standards provide a carefully considered approach that can, with minor modifications, be applied for non-spatial data. With 334 metadata elements defined in the current set, the FGDC metadata standards are the most comprehensive Federal metadata standards. Some FGDC metadata elements are mandatory; many more are "mandatory if applicable" or optional. Some fields may be completed with any-typed text ("free text"). Other fields have specifically enumerated values or require index terms to be drawn from an explicit thesaurus; this improves machine readability and searchability of these records.

> The FGDC framework does not provide enough information about data sets for all applications. Geologists, for example, may require specific, keyword-searchable information about the types of rock strata that might otherwise be described in a free text field. Biologists might need specific information on species or habitat associations. And stewards of former DOE sites might require specific information on special nuclear materials issues, relevance to litigation, or other issues. The FGDC approach includes a provision to create Supplemental Profiles to be used in conjunction with the existing metadata standards. Rather than re-defining existing elements, this process seeks to narrow the options for filling in the existing data elements to assure that the information that is entered is sufficiently specific, and adding additional user-defined data elements as appropriate to capture the information content of the data set.

In all the recommendations for long-term data archiving of geospatial data, geospatial metadata play a prominent role in the success of archiving data for the long-term.

However, geospatial metadata are expensive to collect, the tools are not necessarily easy to use, and once a set of metadata is compiled, there is no guarantee that it will be maintained or that it will be able to be put on an NSDI hub or clearingouse.

Other Federal Agencies

In an effort to understand the complexity of the spatial data archiving problem, three other federal government agencies with responsibilities for large amounts of spatial data were contacted: the USGS, EROS Data Center; NOAA, and the Federal Emergency Management Agency (FEMA). These agencies were contacted for the following reasons: they had an existing working group or process investigating long-term spatial data archiving, they have processes in place to migrate large amounts of spatial data, or they have the legal responsibly for large amounts of spatial data.

To obtain information about spatial data archiving at these agencies, a questionnaire was developed. The purpose of the questionnaire was to ask each agency in a standard way, for information about their data archiving process and for specific information about the archive. The questions were developed from research materials and discussions with recorded information professionals.

Contacts were developed within each of the three agencies, and a series of phone calls to the main points of contacts were made to determine if they had an interest in filling out the questionnaire. Then the questionnaire was sent to the agencies via email. The information presented below was derived from the phone conversations and the answers presented in the questionnaire. The questionnaire is in Appendix A and the list of contacts and responses is in Appendix B.

USGS, EROS Data Center

In 1992, congress directed the Department of the Interior to establish a permanent government archive containing satellite remote sensing data of the Earth's land surface and to make these data easily accessible and readily available for study. This collection is known legally as the National Satellite Land Remote Sensing Data Archive (NSLRSDA). It was established by Public Law 102-555 to be a comprehensive, permanent and impartial record of the planet's land surface derived from almost 40 years of satellite remote sensing.

I spoke with two people involved with the NSLRSDA, Ms. Amy Budge from the Earth Data Analysis Center at the University of New Mexico, and Mr. John Faundeen. Ms. Budge is on the NSLRDA Advisory Committee and Mr. Faundeen is the Chief of Data Management at the EROS Data Center.

The following detailed information about the NSLRSDA is taken from the archive's web site (http://edc.usgs.gov/programs/NSLRSDA.html):

> Over the past three decades the Nation has invested more than $3 billion to acquire and distribute data worldwide from the Landsat series of satellites -- more than 120,000 gigabytes of which are held at the EROS Data Center. This collection from Landsats 1 through 5, including image data from both the Thematic Mapper (TM) and Multispectral Scanner (MSS) sensors forms the core of the national archive but does not complete it.
>
> The archive includes more than 12,000 gigabytes of data from the Advanced Very High Resolution Radiometer (AVHRR) carried aboard National Oceanic Atmospheric Administration's polar orbiting weather satellites and more than 880,000 declassified intelligence satellite photographs.
>
> In addition to these data, the planned archive holdings in 2001 will include data from:
>
> - Landsat 7
> - NASA's MODIS instrument, part of the Mission to Planet Earth's Earth Observing System
> - ASTER, a cooperative effort between NASA and Japan's Ministry of International Trade and Industry
> - The Shuttle Radar Topography Mission (SRTM), a joint venture of NASA and the National Imagery and Mapping Agency
> - LightSAR, a NASA synthetic-aperture radar instrument
> - NASA's Small Spacecraft Technology Initiative, or SSTI

The total holdings by the year 2005 will come to some 2,400,000 gigabytes of data.

To develop this archive, a NSLRDA advisory committee was established to:

- Assist in defining and accomplishing the NSLRSDA's archive and access goals to carry out the requirements of the Land Remote Sensing Act,
- Advise the USGS EROS Data Center (USGS/EDC) on priorities of the NSLRSDA's tasks and
- Provide interdisciplinary guidance and serve as a resource to the USGS/EDC on issues of archiving, data management, science, policy, and public-private partnerships.

This advisory committee has its own web site (http://edcwww.cr.usgs.gov/programs/nslrsda/advisory/index.html) that discusses its charter, justification, meeting minutes, schedule, and members. The advisory committee is unique, in that it is made up of 15 members:

- Two from academia: one laboratory researcher-data user and one classroom educator-data user.

- Four from government: one federal data user, one state data user, one local data user, and one science archivist.

- Four from industry: one data management technologist, one licensed data provider, one value-added or other data provider, and one end user.

- Five others: one non-affiliated individual at-large, one non-governmental organization, one international non-U.S. representative, two at-large from any sector.

Ms. Amy Budge has been on the advisory committee for two years. The advisory committee has been busy working on a series of papers: Justification statement, recommendations for the archive, archive policy white paper, definitions, and a white paper on restricted data. All these papers are available on the web at: http://edcwww.cr.usgs.gov/programs/nslrsda

Early on, the advisory committee realized it needed a way to limit the amount of satellite data that would be archived, it worked on what is called a “Data Sieve” to determine which sets of spatial data would need to be archived. The committee worked on this for several months, believing that it had a good screen. The screen was then tested on several data sets, and almost all data sets fit through the initial screen. Therefore, the committee is working to refine the screen.

In reviewing the reply from Mr. Faundeen to the questionnaire, here are a few pertinent comments:

- The NSLRSDA has a working advisory committee to create guidance and policy for the archive.
- The NSLRSDA works with NOAA, in the sense that each of the archives has copies of the other archive, and was established to meet National Archive and Records Association (NARA) requirements for off –site data storage.
- The NSLRSDA annual archive budget exceeds $4 million annually, and it is anticipating several more million in the near future.
- It is legally bound to archive all data.
- It has an active data migration program, and it currently stores the data in the original data format.
- It uses the SDTS as more of a data distribution format rather than as an archive format
- It uses the FDGC metadata format for recording metadata and as a tool for indexing the data.

NOAA

NOAA is the agency that oversees the National Climatic Data Center (NCDC). NCDC produces numerous climate publications and responds to data requests from all over the world. Their web site is http://lwf.ncdc.noaa.gov/oa/ncdc.html.

Some interesting facts about this archive from the web site:

- The Center has more than 150 years of data on hand with 55 gigabytes of new information added each day--that is equivalent to 18 million pages a day.
- NCDC archives 99 percent of all NOAA data, including over 320 million paper records; 2.5 million microfiche records; over 500,000 tape cartridges/magnetic tapes, and has satellite weather images back to 1960. NCDC annually publishes over 1.2 million copies of climate publications that are sent to individual users and 33,000 subscribers. NCDC maintains over 500 digital data sets to respond to over 170,000 requests each year.
- NCDC supports many forms of data and information dissemination such as paper copies of original records, publications, atlases, computer printouts, microfiche, microfilm, movie loops, photographs, magnetic tape, floppy disks, CD-ROM, electronic mail, on-line dial-up, telephone, facsimile, and personal visit.
- By 2003, the collection will grow by a factor of 15 to approximately 2,000,000 gigabytes of data.

Mr. Ken Davidson, from NOAA's archive data management division, responded to the questionnaire. Below are some of the important points about the NCDC archive:

- It has a formal working group working on spatial data archiving policy.
- It spends about $30 K a year on the formal archive policy development and has approximately two staff years of person time devoted to the project.
- The data from the archive are being used –it has more than 3 million downloads from the web site each year.
- The archive is keeping a "technology archive" of old tape readers and equipment in order to access older data.
- It has an annual data migration plan, and data migration never stops.
- It uses centralized, distributed and multi-node architecture to manage 1.1 petabytes of data.
- It is required to use the FDGC metadata standard and uses it to help index their data.

FEMA

I contacted Mr. Ed Corvi, Team Lead of the GIS and Internet Development Team at FEMA. We had a short discussion on spatial data archiving. FEMA as an agency has been using GIS for disaster management for a long time, however, FEMA only now is looking into enterprise-wide GIS (E-GIS), and spatial data archiving will be part of that E-GIS solution. It currently archives any spatial data associated with disasters with the archive of the disaster data. It develops metadata for the data sets it uses and creates, and it will be looking into the development of more formal procedures for archiving spatial data sets as plans for the E-GIS evolve within the agency.

Conclusions

For DOE's LTS program, a vigorous analysis of the intent of archiving the geospatial data for long-term stewardship will have to be performed. Will the archive be strictly for data preservation across generations, or will it be to provide access to the data for the long-term or some combination of both? This analysis will be fundamental in determining the long-term program plan and long-term costs of the geospatial data archiving process.

To further aid the discussion of geospatial data archiving for LTS, the following list of issues for discussion has been developed:

1. There will be no single grand solution to the spatial data archiving problem. There will need to be a mixture of strategies suitable for different kinds of data.
2. An overall strategy for long-term spatial data archiving will need to be developed. Following the examples from the USGS, NOAA and FEMA, they either have, or are developing, a working group to specifically address data archiving of spatial data.
3. Spatial data archiving is not cheap and has long-term maintenance costs -- hardware, software, and expertise. Commitments to long-term funding for archiving of this nature are necessary for the continuity of the data. Otherwise there is a very large risk of loosing much of the data from neglect.
4. Resolution of the digital data archiving problem is not just technological. It involves management decisions, and policy decisions. Hard decisions will need to be made, and O'Connor (1996) noted:

   > There is a growing awareness that technological development will not be sufficient – many of the challenges involve choices and decisions – these are at the heart, management issues that will have to be addressed with the aid of technology.

5. Technology issues: Above and beyond the issues of technological obsolescence discussed, other technology issues will become part of the discussion:

   - ESRI's change to ArcGIS architecture will have a huge impact on DOE's LTS program. Much of the existing GIS data across the DOE complex may never be migrated to the new architecture and, without documentation and preservation could be lost forever.
   - Storage volume is becoming cheaper -- there will be the tendency for data managers to just purchase more storage space rather than deal with the issues moving data into an archive.

6. For DOE's LTS program, the issue of spatial data metadata will become crucial. All of DOE's facilities are required to capture some form of metadata. Currently,

metadata are in a variety of formats, and it is unknown if they could be served to a metadata server within the DOE Complex. Metadata will become a key to the long-term maintenance of these data sets

7. Data formats and compression. DOE will need to develop some guidance on the preferred data formats. This is not to suggest a single data format for all sites. Some type of recommendation will need to be given on the acceptable formats and data compression. Most of the literature on digital data preservation does not recommend data compression of any type, but at some DOE facilities, data compression is used regularly, especially for image compression.

8. Short of any policy for geospatial data archiving, "Keeping the data alive" (i.e., keeping a GIS system active and migrating the data and system forward in time) may be the best temporary solution. Until some type of guidance is developed, there are too many unknowns for "closing down and archiving" an existing GIS system.

Suggested Path Forward

Spatial data archiving will be a large issue for DOE as it moves forward in the development of the policies for LTS. One suggested path forward would be to form a small working committee of GIS practitioners from DOE sites and field offices and information management specialists to start looking at and developing some initial guidance on the spatial data archiving issue. By using the USGS and NOAA archives as examples, and the white papers that these agencies have developed concerning these archives, DOE could save some time and effort in developing geospatial data archiving policies. The goal of this small working group would be to review existing spatial data archives and to develop a set of guidelines for the GIS practitioners at each of the DOE facilities.

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.

Appendix A : Agency Questionnaire

Process

1. Is your agency using a formalized process to investigate long-term spatial data storage or data archiving?
2. If so, what is that process (working group, task groups, etc.)?
3. Is your agency interacting with other federal agencies or groups (private sector?) working on this issue?
4. Has your organization published anything on the issues around long-term spatial data storage or archiving? If so, may we have copies?
5. How much is this process costing your agency or group?
6. How much effort is this taking within your agency or group?
7. Has your agency addressed the long-term funding issues for spatial data archiving?

User Issues:

1. Who really needs this spatial data (who is the intended user group of the spatial data sets in the archives)?
2. How are you deciding what spatial data to keep?
3. What are the purposes for keeping this data? How do you intend to use the data? (general science based on change detection, legal reasons, basic reference, etc)?
4. Do you really think this archived data will be used? If so, how did you make this determination? Or, is it more for "just in case…."?
5. Does the data really need to be digital?
6. How did you consider users who may use this data in the future (generations from now?)

Access Issues:

1. Is the data "live and on line" for immediate access? Are you planning on using the web?
2. Are the data on tape that would have to be requested, loaded and then viewed?
3. Are the data indexed in any way?
4. Are the data stored by date, sensor type, geography, theme or project or in a combination?

Data Retention:

1. How are you determining how long you need to keep the spatial data?
2. Are you basing retention on laws, policies, specific guidance or user's needs? If so, which guidance are you following?
3. Is the guidance reflected in your retention schedule?
4. What are the drivers requiring you to keep digital data?
5. What is the final disposition for these data (will it be sent to another agency, NARA, etc.)?

Media:

1. Media -- how are you determining what media will be used to store spatial data for the long-term?

Migration:

1. Migration issues - who will be responsible for data migration in the future as technology changes?
2. Who will pay for data migration in the future?
3. Is your agency intending to keep a "technology archive" of equipment -- i.e. old tape readers, drives, etc?
4. Migration issues - how are you handling the current migration of spatial data forward from the old 9-track tapes?

Platforms:

1. Which platform are you using for your archive (UNIX, NT, other)
2. Is the data "on line" or retrievable from tape?
3. What is the system architecture of the archive (centralized distributed, multi-node, etc)?

(Do you care if and how they are backing up the data?)

Data Compression:

1. Are you using data compression software? If so, what?
2. Why did you choose this compression software?

Data Format:

1. What data format are you storing your spatial data in?
2. Are you converting it to a basic format (such as the SDTS?)
3. Are you keeping the data in its native format?
4. If you are planning to convert the data, do you have a quality assurance process?

Spatial Data Metadata:

1. Are you using spatial data Metadata as an index to your spatial data archive?
2. If so, which metadata tool(s) are you using?
3. Have you consulted with any information specialists regarding metadata standards?
4. Do you have written metadata documentation?

Appendix B: Agency Contacts and Responses

1. United States Geological Survey, EROS Data Center
   Mr. John Faundeen
   Email: faundeen@usgs.gov

   Website for more information on the National Satellite Land Remote Sensing Data Archive: http://edcwww.cr.usgs.gov/programs/NSLRSDA.html

2. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA)
   Mr. Ken Davidson
   Email: kenneth.d.davidson@noaa.gov

   Website for more information on the National Climatic Data Center: http://www.ncdc.noaa.gov/

3. Federal Emergency Management Agency (FEMA)
   Mr. Ed Corvi
   Email: ed.corvi@fema.gov

EROS Data Center Response

Denise,

My answers are below your questions at the end of this email. We would like to request copies of the final report.

Thank you,

John Faundeen
Chief, Data Management
U.S. Geological Survey
EROS Data Center
Sioux Falls, SD 57198 USA

Tel: 605-594-6092
Fax: 707-222-0223

Questions to consider for long-term spatial data archiving

Process

1. Is your agency using a formalized process to investigate long term spatial data storage or data archiving?

We periodically fund internal trade studies to determine which next-generation media we should evolve to. All digital media is migrated within a 10 year cycle due to media and hardare obsolescence. In the past, we also sought out National Media Labs recommendations. They disbanded this Spring.

2. If so, what is that process (working group, task groups, etc)?

We have an Archive Program within our facility that handles all aspects of data management. Contractors carry out the actual work.

3. Is your agency interacting with other federal agencies or groups' (private sector?) working on this issue?

Beyond the National Media Labs mentioned above, we are active in the international Committee on Earth Observation Satellites which includes data management topics. We have also begun a reciprocal off-site archive arrangement with the NOAA National Climatic Data Center in Ashville, SC. This arrangement, begun this fiscal year, has each Agency store the other Agencies' data in a controlled facility to meet the NARA offsite

requirement. To date, NOAA has shipped 54 pallets of data to us and we have shipped 22 pallets to them. This is an ongoing arrangement set up by a Letter of Agreement between our two Centers.

4. Has your organization published anything on the issues around long term spatial data storage or archiving? If so, may we have copies?

See our online, searchable, pubs area (http://edc.usgs.gov/content_pub.html) and my attachment of citations that deal with the Archive and access tothe Archive topics.

5. How much is this process costing your agency or group?

Our annual Archive Program budget is >$4M annual. We are planning on seeking a Congressional initiative for several more millions in the near future.

6. How much effort is this taking within your agency or group?

Considerable. See above response.

7. Has your agency addressed the long-term funding issues for spatial data archiving?

We were successful in one Congressional initiative to address long-term issues, but the new data rates may swamp us if we do not receive additional base monies.

<u>User issues:</u>

1. Who really needs this spatial data (who is the intended user group of the spatial data sets in the archives)?

Environmental organizations/researchers, geologists, land use researchers. See our online pubs for the diversity of applications for our land-based data.

2. How are you deciding what spatial data to keep?

Question of the day......We have not thrown anything away yet. Personally, I feel that this is a scence issue that only the National Academy of Sciences or NSF should deal with. Not me as data manager.

3. What are the purposes for keeping this data? How do you intend to use the data? (general science based on change detection, legal reasons, basic reference, etc)?

See No. 1 above, but note that legal uses do abound.

4. Do you really think this archived data will be used? If so, how did

you make this determination? Or, is it more for "just incase....."?

Yes, I think the data will be used. Regardless, the Land Remote Sensing Policy Act of 1992, the National Space Policy of 1996, the U.S. Global Change Research Program all have language that directs DOI > USGS to maintain this archive.

5. Does the data really need to be digital?

Interesting question. I know NIMA plans to produce film products from some of its data received from NTM and to then delete the original source, digital data strategizing that they can always scan the film to get back to a digital source on a demand basis. We plan to continue archiving the data as it comes to us, i.e. either on film (DeClass I) or digital (Landsat).

6. How did you consider users who may use this data in the future (generations from now?)

Preservation for future generations has been part of our mindset for quite some time.

<u>**Access issues:**</u>

1. Is the data "live and on line" for immediate access? Are you planning on using the web?

The metadata, and digital browse representations are online typically within one day of acquisition. We are investigating web mapping mechanisms to serve the data online, upto 1:1 resolution. Couple of hurdle there: A) Most of the data is stored offline and B) The data volumes may exceed bandwidth realities. We plan on providing some satellite data to users this way in FY02, though.

2. Is the data on tape that would have to be requested, loaded and then viewed?

Currently, users utilize a system like http://earthexplorer.usgs.gov to determine what is available and then look at preview, sub-sampled images to make order decisions. The 1:1 resolution data is then produced offline.

3. Is the data indexed in any way?

Yes. See the URL listed in No. 2 above.

4. Is the data stored by date, sensor type, geography, theme or project or in a combination?

Searchable by all of those means, not stored in a database that way. Spatial and temporal elements are the first critical means most of our

users cut through our holdings with.

<u>Data Retention</u>
1. How are you determining how long you need to keep the spatial data?

We consider anything in the NSLRSDA to be forever.

2. Are you basing retention on laws, policies, specific guidance or user's needs? If so, which guidance are you following?

Yes to all, but the laws and policies quoted above are the critical drivers.

3. Is the guidance reflected in your retention schedule?

Yes.

4. What are the drivers requiring you to keep digital data?

Same. Laws, policies, Agency direction.

5. What is the final disposition for this data (will it be sent to another agency, NARA, etc.)?

The NSLRSDA data will always reside here. Other data, such aerial photography, will probably migrate to NARA.

<u>Media</u>
1. Media -- how are you determining what media will be used to store spatial data for the long-term?

See No. 1 under Process above.

<u>Migration</u>
1. Migration issues - who will be responsible for data migration in the future as technology changes?

The Archive Program. We are soliciting a government position for an Archivist to assist us in this responsibility. The announcement closes tomorrow.

2. Who will pay for data migration in the future?

Hopefully, Congressional dollars.

3. Is your agency intending to keep a "technology archive" of equipment
-- i.e. old tape readers, drives, etc?

Only to support the data we currently have, i.e. not for media we have migrated away from.

4. Migration issues - how are you handling the current migration of spatial data forward from the old 9-track tapes?

Transcribing them to DLTs today and finishing up a trade study to point the way for our next generation media.

<u>**Platforms**</u>

1. Which platform are you using for your archive (UNIX, NT, other)?

Unix generally, but NT front-ends are starting to be utilzed, too.

2. Is the data "on line" or retrievable from tape?

From offline tape.

3. What is the system architecture of the archive (centralized distributed, multi-node, etc)?
(Do you care if and how they are backing up the data?)

Offline in an environmentally controlled archive. Plan to have third copies stored offsite (mentioned NCDC above).

<u>**Data Compression**</u>

1. Are you using data compression software? If so, what?

Lossless only, like unix compression.

2. Why did you choose this compression software?

Lossless is a requirement.

<u>**Data Format**</u>

1. What data format are you storing your spatial data in?

Computer-compatiable now. Originally stored the analog as is. Now converting for easier reading and future migrations.

2. Are you converting it to a basic format (such as the SDTS?)

No. We treat that as a distribution element.

3. Are you keeping the data in its native format?

Lowest-level possible is our goal. This keeps us flexible in product offerings.

4. If you are planning to convert the data, do you have a quality assurance process?

Yes. Each transcription process has this. We will have four separate migration operations ongoing next FY.

<u>Spatial Data Metadata</u>

1. Are you using spatial data Metadata as an index to your spatial data archive?

Yes. It is critical to us and to our users.

2. If so, which metadata tool(s) are you using?

Oracle SQL. ESRI SDE. Investigating CubeWerx and Microsoft SQL among others.

3. Have you consulted with any information specialists regarding metadata standards?

We support NSDI / FGDC standards and spend a fair amount addressing those needs annually.

4. Do you have written metadata documentation?

Our Information Program has for years indicated that we must support FGDC, ISO, and OGC standards.

NOAA Response

From: Kenneth D. Davidson [Kenneth.D.Davidson@noaa.gov]
Sent: Tuesday, September 04, 2001 1:07 PM
To: Bleakly Denise R.
Subject: Re: Long-term Spatial Data Archiving

Questions for consideration on Spatial Data Archiving

Process

1. Is your agency using a formalized process to investigate long-term spatial data storage or data archiving?

YES

2. If so, what is that process (working group, task groups, etc)?

A FORMAL WORKING GROUP

3. Is your agency interacting with other federal agencies or groups' (private sector?) working on this issue?

YES

4. Has your organization published anything on the issues around long-term spatial data storage or archiving? If so, may we have copies?

YES THEY HAVE BEEN IN SEVERAL JOURNALS. PLEASE SORT THROUGH THEM LOOKING FOR NCDC.

5. How much is this process costing your agency or group?

$30K PER YEAR

6. How much effort is this taking within your agency or group?

2 STAFF YEARS

7. Has your agency addressed the long-term funding issues for spatial data archiving?

WE HAVE TRIED FOR SEVERAL YEARS, BUT TO NO AVAIL

User issues:

1. Who really needs this spatial data (who is the intended user group of the spatial data sets in the archives)?

ALL USERS, FROM LAWYERS TO RESEARCHERS.

2. How are you deciding what spatial data to keep?

A WORKING GROUP ADVISES THE DIRECTOR, WHO MAKES THE FINAL DECISION

3. What are the purposes for keeping this data?
WE ARE LEGALLY REQUIRED TO MAINTAIN THESE DATA

3a. How do you intend to use the data? (general science based on change detection, legal reasons, basic reference, etc)?

YES ALL OF THESE

4. Do you really think this archived data will be used? If so, how did you make this determination? Or, is it more for "just incase....."?

YES IT WILL AND IS BEING USED. WE HAVE OVER 3,000,000 DOWNLOADS FROM OUR WEB SITE EACH YEAR. OVER 300,000 REQEUSTS THAT ARE SERVICED OFF-LINE. MOST OF THESE ARE PAYING DATA USERS

5. Does the data really need to be digital?

IT MUST BE, WE COULD NOT SERVICE THE USERS IF IT WERE NOT.

6. How did you consider users who may use this data in the future (generations from now?)

WE TRY TO DO THAT, BUT IT IS DIFFICULT

<u>Access issues:</u>

1. Is the data "live and on line" for immediate access? Are you planning on using the web?

YES

2. Is the data on tape that would have to be requested, loaded and then viewed?

SOME OF THE LESS FREQUENTLY USED DATA ARE ON TAPE OR ON TAPE IN A MASS STORAGE SYSTEM.

3. Is the data indexed in any way?

YES

4. Is the data stored by date, sensor type, geography, theme or project or in a combination?

GENERALLY IT IS MAINTAINED IN A STATION AND SYNOPTIC SORT.

<u>Data Retention</u>

1. How are you determining how long you need to keep the spatial data?

LEGALLY WE ARE REQUIRED TO MAINTAIN THE DATA ACCORDING TO OUR DISPOSITION PLANS THAT HAVE BEEN AGREED TO BY THE NATIONAL ARCHIVIST.

2. Are you basing retention on laws, policies, specific guidance or user's needs? If so, which guidance are you following?

LAWS

3. Is the guidance reflected in your retention schedule?

YES

4. What are the drivers requiring you to keep digital data?

WE COULD NOT SERVICE THE USERS IF IT WERE IN HARDCOPY OR OFF-LINE ONLY.

5. What is the final disposition for this data (will it be sent to another agency, NARA, etc.)?

DEPENDS ON THE DATA TYPE. SOME ARE SENT TO THE NATIONAL RECORDS CENTER IN EAST POINT, GEORGIA, OTHERS CAN BE DESTROYED

<u>Media</u>

1. Media -- how are you determining what media will be used to store spatial data for the long-term?

COST AND APPROVAL BY THE NATIONAL ARCHIVIST

Migration

1. Migration issues - who will be responsible for data migration in the future as technology changes?

OUR AGENCY

2. Who will pay for data migration in the future?

TAXPAYERS FROM APPROPRIATED FUNDS

3. Is your agency intending to keep a "technology archive" of equipment -- i.e. old tape readers, drives, etc?

YES

4. Migration issues - how are you handling the current migration of spatial data forward from the old 9-track tapes?

WE HAVE AN ANNUAL PLAN AND MIGRATION NEVER STOPS.

Platforms

Which platform are you using for your archive (UNIX, NT, other)?
YES

Is the data "on line" or retrievable from tape?

BOTH

3. What is the system architecture of the archive (centralized distributed, multi-node, etc)?

YES, ALL TYPES ARE USED BY THE CENTER. THE CENTER HAS OVER 1.1 PETABYTES OF DATA
(Do you care if and how they are backing up the data?)

Data Compression
Are you using data compression software? If so, what?

YES STANDARD SOFTWARE

Why did you choose this compression software?

REQUIRED TO MOVE DATA TO THE MASS STORE

Data Format

1. What data format are you storing your spatial data in?
2. Are you converting it to a basic format (such as the SDTS?)

Are you keeping the data in its native format?

YES

4. If you are planning to convert the data, do you have a quality assurance process?

YES

Spatial Data Metadata

1. Are you using spatial data Metadata as an index to your spatial data archive?

YES

2. If so, which metadata tool(s) are you using?
3. Have you consulted with any information specialists regarding metadata standards?

YES

Do you have written metadata documentation?

YES

Distribution

	External Copies
10	Randy D. Lee Idaho National Engineering and Environmental Laboratoy P.O. Box 1625 Idaho Falls, ID 83415-2213
5	Karin Brown DOE-ID Ops Office M/s 1240 850 Energy Drive Idaho Falls, Idaho 83401
5	Susan Hargrove Office of the Chief Information Officer U.S. Department of Energy, IM-50 1000 Independence Ave. S.W. Rm. 8H-089 Washington, D.C. 20585
1	John Stewart Office of Long Term Stewardship, EM-51 U.S. Department of Energy, Headquarters Germantown 19901 Germantown Road Germantown, MD 20874-1290
1	Michael Barainca U.S. Department of Energy, Headquarters, EM-51 Germantown 19901 Germantown Road Germantown, MD 20874-1290
1	David Geiser Office of Long Term Stewardship, EM-51 U.S. Department of Energy, Headquarters Forrestal 1000 Independence Avenue, S.W. Washington, DC 20585
1	Susan Frey Records Management Division, IM-11 U.S. Department of Energy, Headquarters Germantown 19901 Germantown Road Germantown, MD 20874-1290
1	Cathy Marciante Information Resources Management Division U.S. Department of Energy Oak Ridge Operations Office 200 Administration Road Oak Ridge, TN 37831

2	Deborah Couchman-Griswold U.S. Department of Energy Albuquerque Operations Office SC-1 Pennsylvania & H Street Kirtland Air Force Base Albuquerque, NM 87116
1	Joe Estrada U.S. Department of Energy Kirtland Area Office Pennsylvania & H Street Albuquerque, NM 87185-5400 Al Guber Remote Sensing Laboratory P.O. Box 98521 Las Vegas, NV 98193-8521 Steve Livingstone ICF Consulting 9300 Lee Highway Fairfax, VA 22031
1	Bob Hegner ICF Consulting 9300 Lee Highway Fairfax, VA 22031
1	Robert Andrew ICF Consulting 9300 Lee Highway Fairfax, VA 22031
1	Greg Csullog WMDB/WMRA Programme Officer International Atomic Energy Agency Room A2656 Wagramerstrasse 5, P.O. Box 100 A-1400, Vienna, Austria
1	Eberhard Falck WMDB/WMRA Programme Officer International Atomic Energy Agency Room A2656 Wagramerstrasse 5, P.O. Box 100 A-1400, Vienna, Austria
1	Mr. Stewart Brand The Long Now Foundation P.O. Box 29462 Presidio of San Francisco San Francisco, CA 94129-0462

2		John Faundeen Chief, Data Management U.S. Geological Survey EROS Data Center Sioux Falls, SD 57198 USA
1		Sharon LeDuc National Climatic Data Center 151 Patton Ave Ashville, NC 28801
1		Mr. Ed Corvi Team Leader GIS and Internet Development Team 500 C. Street SW Room 226 Washington D.C., 20427
		Internal Sandia Copies
6	MS-0763	Denise Bleakly, 6135
1	MS -0612	Linda Cusimano, 9612-1
1	MS-0612	Jean Ann Plummer, 9612-1
1	MS-0612	Barbara Staley, 9612-1
10	MS-0939	Peggy Warner, 9612
1	MS-0719	Susan Howarth, 6131
1	MS-0719	Warren Cox, 6131
1	MS-1089	Dick Fate, 6135
1	MS-0763	Steve Ratheal, 5853
		Housekeeping copies
1	MS-9018	Central Technical Files, 8945-1
2	MS 0899	Technical Library, 9616
1	MS 0612	Review & Approval Desk, 09612 for DOE/OSTI