



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

**4th Annual Long Term Stewardship Workshop
August 1, 2001**

**Denise Bleakly
Sandia National Laboratories**



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.





Introduction

As DOE prepares to enter into Long-Term Environmental Stewardship (LTES), decisions will be made concerning the long-term care and maintenance of *digital* geo-spatial data.

Each of the DOE sites that will be entering into LTES has some form of digital geo-spatial data

This presentation will briefly discuss issues around the long-term management of digital geo-spatial data



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**
- **Over the last 30 years there has been a shift from primarily paper records to electronic records**



Issues for Digital Data Preservation

- **Media**

- **Digital media can be fragile and have a limited life-span (less chemically stable than paper)**
- **Is machine-dependent - machines must read the data**
- **Is totally system-dependent for retrieval of information (both hardware and software)**
- **Data compression techniques make information retrieval vulnerable to large losses from small data errors**
- **Failure of media is unpredictable and sudden and can result in total loss of data**



Issues Cont'd

- **Technological Obsolescence**
 - Every 2-5 years new devices, processes and software are replacing the products and methods used to record, store and retrieve digital information (remember 8-track tapes?)
 - Lack of compatibility between hardware and software platforms
 - Most new software is not “backwardly compatible”



Issues Cont'd

- **Data Refreshing**
 - Copies data from older medium to newer medium
 - Does not guarantee that the information will be usable with new software
- **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
 - Tries to retain the ability to display, retrieve, manipulate and use the information



Issues Cont'd

- **Emulation**
 - building software that makes other software act as if it was something else
 - Is now being tested for reading early digital text formats (remember Wang word processors?)
- **Long-term Costs**
 - **Preservation of digital data is expensive**
 - Knowing what to migrate - subject matter experts and information professionals
 - Migration of the data itself
 - Building and maintaining indexes to archived information
 - Hardware/software



Greatest Fear of Archivists

- **Owners or custodians who can no longer bear the expense and difficulty will deliberately or inadvertently, through a simple failure to act, destroy information without regard for future use**

Examples

- **Early satellite data**
- **Canadian Land Information System**
- **Census data**



Data Archiving versus Information Preservation

- **Digital data can be well-archived but not well-preserved:**
 - Data copying can ensure that the information is archived on the newest media, but loose the integrity of the information
 - Data archiving may facilitate preservation but not ensure it.
- **Data preservation is concerned with the long term retrievability and use of information within its original data context.**



What Makes Geo-Spatial Data Unique?

- **Geo-spatial data represent many facets of phenomena on the earth and are stored as points, lines, polygons, regions, volumes, grids**
- **Geo-spatial data stored in a GIS usually have relationships between objects stored as part of the data structure**
- **Geo-spatial data are multi-scaled and have multi-resolutions**
- **The power of geo-spatial data is the ability to derive new data from relationships between data layers**
- **Geo-spatial data can be both current and historical**
- **Geo-spatial data can be in the form of aerial photos, maps, surveys, GPS data, etc.**



Why is Geo-Spatial Data Difficult to Archive?

- Many different software platforms: ESRI, Intergraph, AutoCad, MapInfo, etc.
- Many different data formats: CAD, GIS, geo-spatial models, relational databases
- Volumes of data - many geo-spatial systems store *terabytes* of information
- Site-specific spatial data organization
- No consistent use of spatial data metadata for indexing and cataloging



Existing Guidance

- **The Federal Geographic Data Committee has created the fact sheet**
“Managing Historical Geospatial Data Records: A Guide for Federal Agencies”
 - <http://www.fgdc.gov/nara/hdwgfsht.html>
 - **Provides a general overview of Federal Agencies’ responsibilities for the preservation of historical geospatial data, but not “how to”**



Deciding What to Keep

- **What geo-spatial data need to be preserved?**
 - This is a long-term, on going discussion among records management, information management, and archivists.
 - 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 LTES program, this has yet to be defined



Data Storage Versus Data Access

- **Recently spatial data archives have been defined as a way to access geo-spatial data**
 - **There are any number of “Clearing Houses”, “Spatial Data Warehouses”, “Data Archives” on the Internet as a way to access Federally funded geo-spatial data:**
 - **National Satellite Land Remote Sensing Data Archive**
 - <http://edc.usgs.gov/programs/NSLRSDA.html>
 - **National Geospatial Data Clearinghouse**
 - <http://130.11.52.184/>



Data Access Cont'd

- **The use of geo-spatial data archives for data access is very different from “traditional” data archives -- in that it is assumed that the data are “on-line” or live and not stored in an off-line archive**
- **This will have ramifications for DOE’s LTES program.**
- **Will geo-spatial data archiving for LTES mean the use of geo-spatial archives for accessing LTES data?**



Conclusions

- **No single grand solution - there will need to be a mixture of strategies suitable for different kinds of data**
- **Storage volume is becoming cheaper -- this will be less of a problem in the future**
- **However, technology is changing faster than ever!**
 - **ESRI's change to ArcGIS architecture is an example that will have ramifications for DOE's LTES program**



Conclusions Cont'd

- **The use of spatial data metadata may assist in the long-term indexing and accessing of spatial data sets**
- **Storage of information in a format that is independent of the particular hardware and software needed to use it:**
 - **Spatial Data Transfer Standard (SDTS)**
 - **The SDTS has not been uniformly adopted or used**
 - **The new GXML - for geographic data is being proposed by the Open GIS Consortium**



Conclusions Cont'd

- It is recommended **NOT** to use any form of data compression software for data files (again - an issue of data migration forward)
- “Keeping the data alive” (i.e., keeping a GIS system active and migrate the data and system forward in time) may be the best temporary solution for geo-spatial data.
 - There are too many unknowns for “closing down and archiving” a existing GIS system