DATA IN JEOPARDY

Authors: Dr. Kian Fadaie (Canadian Centre for Remote Sensing), Trevor Milne, (Helical Systems Ltd.), Herman P. Varma (Canadian Hydrographic Service), Jennifer Harding, Ronald Macnab and Pierre Gareau (Geological Survey of Canada) Douglas O'Brien (IDON Corporation)

Abstract:

During the last 30 years, there has been a serious neglect of the information base stored within our data holdings.

In the 1800's data was stored in media such as books, which had encapsulated indexes, meta information, versioning and data. The media, paper, was stable and could be opened up a hundred years later and the information could still be retrieved. Due in large part to rapid technology shifts and the lack of appropriate international standards, this is no longer the case.

For the past 30 years, technology shifts have created serious problems in large data repositories. The constant change of storage media from punched cards, to paper tape to 800 BPI magnetic tapes, to cartridges, to magnetic disks, to optical media etc. has resulted in major problems for very large data archives. The hardware and software required to read the old media may not be available in a hundred years, leaving the entire information base in a crisis state.

The technology providers also constantly change application software to store and read data. These changes are sometimes implemented within the space of six months, such as in GIS and Wordprocessing software applications. The software is not always backward compatible due to the encapsulation of proprietary file architectures, algorithms and smart compression. These file architectures, algorithms, compression or otherwise, may not be available if the company should ever go out of business or radically change the application domain.

There is an urgent need to stabilize the storage data structures to some open international standard for long term archival. If not, the large data holdings will be in jeopardy - not in a hundred years but in less then ten years.

A generic storage mechanism, SDS (Self Defining Structure), has been proposed to the ISOTC211 WG1 under the guidance of the Image and Gridded Data Working Group.

This paper addresses problems encountered in the state of the data sets used for Canada's Law of the Sea project and possible solutions using SDS (Self Defining Structures) and distributed archives.

1. Introduction

Most data managers and users of large databases are familiar with situations where vital holdings of scientific and business information would have become irretrievably lost on account of technological and software changes, if not for the major allocations of resources and funds to save the data. A prime example of this was the joint NASA/NSF/NOAA effort in 1990-99 to recover several years' worth of TOVS/AVHRR satellite data from an obsolete computer system. The herculean and expensive effort to save this information resulted in the preservation of a valuable record that documented global warming over a 20-year period. Data loss can be minimized or even prevented with the use of better data archival strategies as designed and promoted by international groups such as the International Standards Organization Technical Committee 211 on Geomatics (ISO TC211) for Image and Gridded Data. Unfortunately, the operating priorities of many world and national data centers force them to concentrate almost exclusively on increasing the power of existing computer infrastructures in order to cope with escalating data flows. They are less able to focus on preparing for the future by seeking and adopting innovative storage techniques and methodologies. A number of anecdotal examples can be cited to demonstrate the significance of data losses which could have been prevented if these types of data strategies had been made available at the national and international standards level.

2. Rapid Data Growth

The rate of data acquisition has experienced exponential growth in recent years. For example, there are programs that deploy a variety of satellite, airborne, and shipborne sensors for the intensive mapping of landforms and seafloor. These programs are collecting vast quantities of data that exceed terabytes (TB 1000 gigabytes) to petabytes (PB 1000 terabytes) a year in range. They produce enormously large scientific data holdings with substantial requirements for archival and retrieval facilities. Current database technologies are not designed to handle extremely large, multidimensional data sets for long term archival. This is due, in large part, to the instability of the RDBMS applications themselves, which are constantly evolving and changing internal structures with each new version. This does not ensure backward compatibility of database applications which are necessary for long term archival. Clearly this has a direct impact on the stability and usability of very large data sets.

In an upheaval as great as the introduction of printing itself, information technology is revolutionizing the concept of record keeping. The current generation of digital records has unique historical significance with the advent of the Internet and other technological breakthroughs. Yet these records are far more fragile than paper, placing the chronicle of this entire period in jeopardy.

Two realities define the current state of affairs:

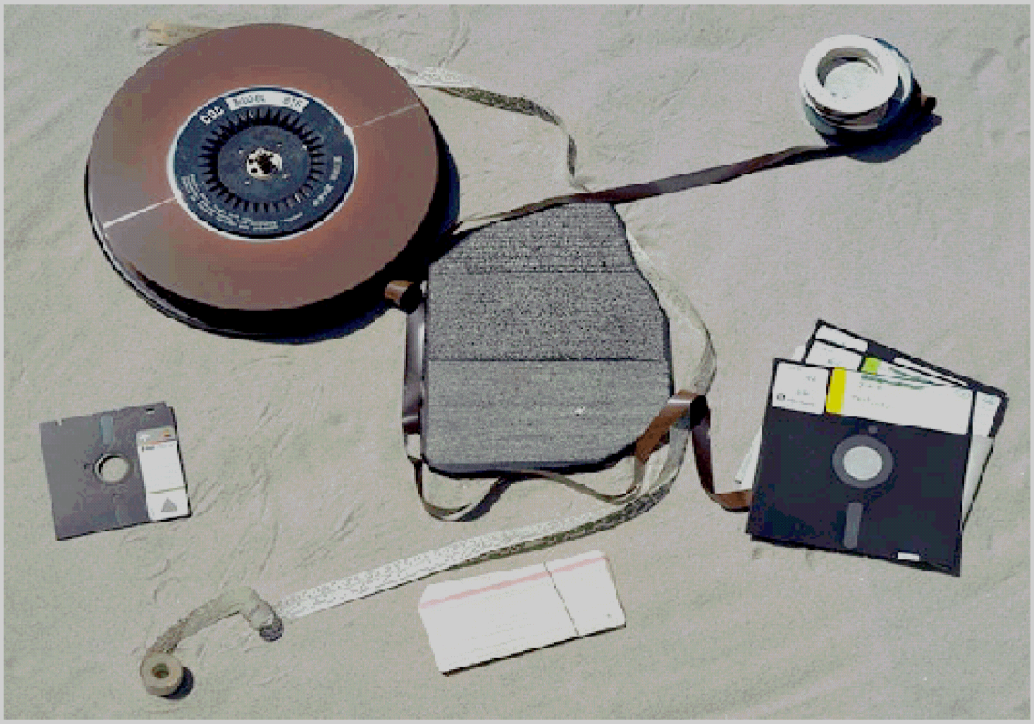
(i) Hardware and software components tend to become obsolete sooner than the storage media, leaving behind massive volumes of data in proprietary legacy systems.

(ii) Some key software systems are evolving without maintaining backward compatibility in terms of device readability, portable file management, network connections and other operability issues.

3. Media and Technology Issues

To understand the basis of the problem, it is necessary to examine the nature of digital storage. Digital information can be stored on any medium in the form of bitstreams. In the past four decades, the types of media for storing these bitstreams have ranged from punched cards to paper tape to magnetic tape to magnetic disks to optical disks. In order to prevent the loss of digital information stored on such media, it has been necessary to copy digital information on a regular basis onto new forms of media, thereby maintaining their longevity and accessibility (Rothenberg 1995). This approach is analogous to preserving text which must be transcribed or reprinted periodically.

The future survivability of digital information depends on an unbroken chain of migrations, which must be frequent enough to prevent the media from becoming physically unreadable or obsolete before they are copied. A single break in this chain can render digital information inaccessible and lead to data loss. Migrations are required as frequently as once every few years to mitigate the transient nature of the media. Eventually, the development of long-lived storage media will make media migration less urgent or frequent. However, frequent changes in software applications, coupled with the lack of encapsulated meta information with the data, continue to pose serious risks to the future survivability of digital information.



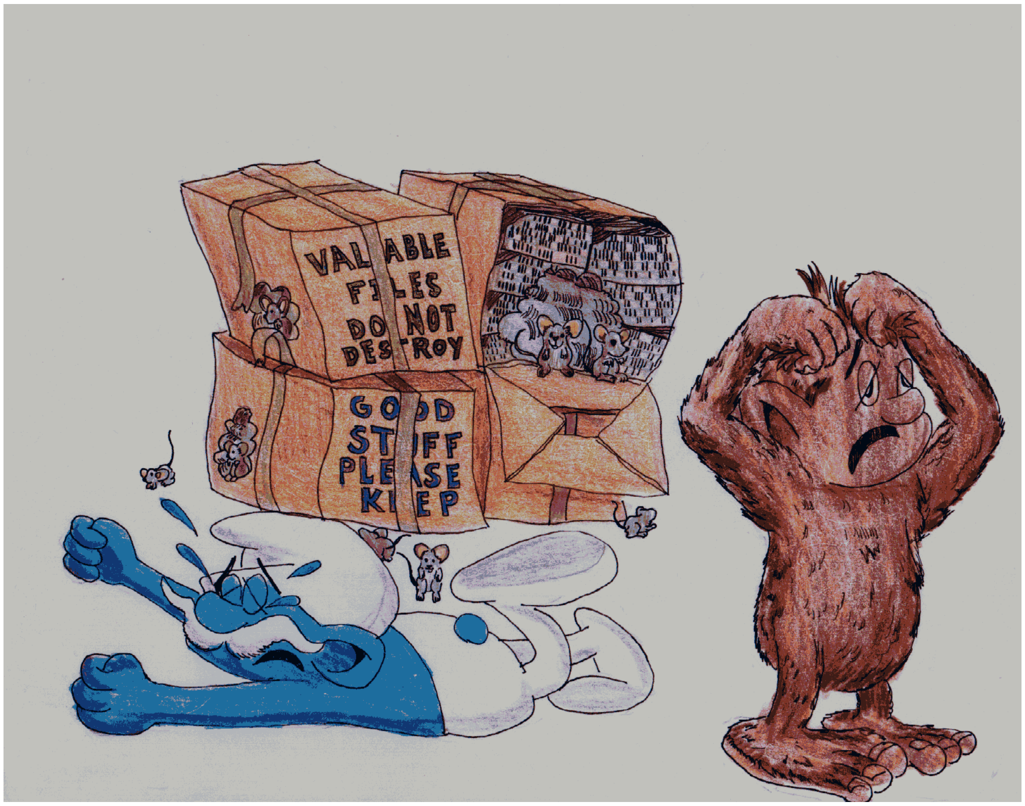
Color photo by Jeff Rothenberg

Photo 1: The Rosetta Stone has far outlasted digital media

In light of these concerns, two strategies have evolved for preserving digital information:

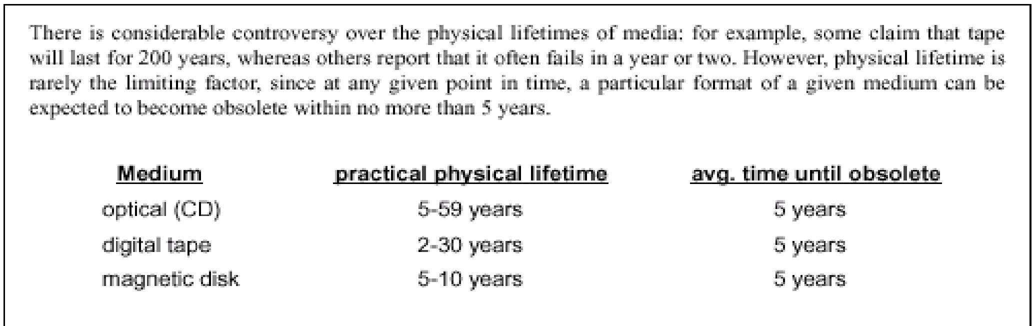
1) Translate records into standard forms that are independent of any computer system.

2) Extend the longevity of computer systems and their original software to maintain the readability of documents.



“Doug,! I thought you told me punched cards were more stable than magnetic tape”

Unfortunately, both strategies have potential shortcomings. Unless digital information is stored in a standard form that encapsulates essential meta information, future usability can be compromised through the loss of specifications of file schema, file structure and the data itself. Moreover, physical storage media is far from eternal. Paper media, being organic, is subject to decay. Magnetic disks and tapes are vulnerable to stray magnetic fields, oxidation, and material decay. CD-ROMs have oxidation problems that cause the digital pits to degrade over time. Gold platters are subject to flow problems that cause annealing of the digital pits. The contents of most digital media, in fact, may disappear long before records that are printed on high quality paper. Even if the media were more long-lived, the information they contain can be rendered obsolete as computer systems and software applications are continually revised to meet escalating demands for



There is considerable controversy over the physical lifetimes of media: for example, some claim that tape will last for 200 years, whereas others report that it often fails in a year or two. However, physical lifetime is rarely the limiting factor, since at any given point in time, a particular format of a given medium can be expected to become obsolete within no more than 5 years.

Medium	practical physical lifetime	avg. time until obsolete
optical (CD)	5-59 years	5 years
digital tape	2-30 years	5 years
magnetic disk	5-10 years	5 years

Figure 2: The medium is a short-lived message

throughput, or for profitability in a highly competitive marketplace.

4. Storage Density, Information Growth, and Data Transfer Rates

Over the past 10 years, storage density on magnetic media has doubled about every 18 months. During the same period, data transfer speeds have only increased at a rate of about 1.3 times and have fallen behind the growth rates of data density by a factor of at least three. Coupled with the growth of media density and cpu performance, data storage requirements have also gone up significantly. According to a recent Computer Technology Review article (March 1998) the total storage at a typical Fortune 1000 site is projected to escalate from just 10 TB (Terabyte, 1000 gigabytes) in 1997 to 1 PB (Petabyte, 1000 Terabytes) by the year 2000, a factor of 100. In the next 5 years, it is anticipated that a large database system of the sort operated by a typical U.S. government agency will have to accept 5 TB per day, to maintain 300 TB on-line (with access time ranging from 15 seconds to 1 minute), and to archive from 15 to 100 PB (Halem 1999).

The information revolution has placed a great demand on the storage industry to develop media that can keep up with exponential growth at constant costs. Recent technology developments have resulted in the ability to store a PB of data in silos for under $1 million (US), which amounts to about $1/GB for the media alone. Right now, this local capacity is large enough to handle most storage demands for all known civilian systems, except for a few large projects such as the Earth Observation System. However, at current transfer rates, the migration of a few PB of data to a new media for preservation could take more than 10 years. Hence, the crux of the crisis:

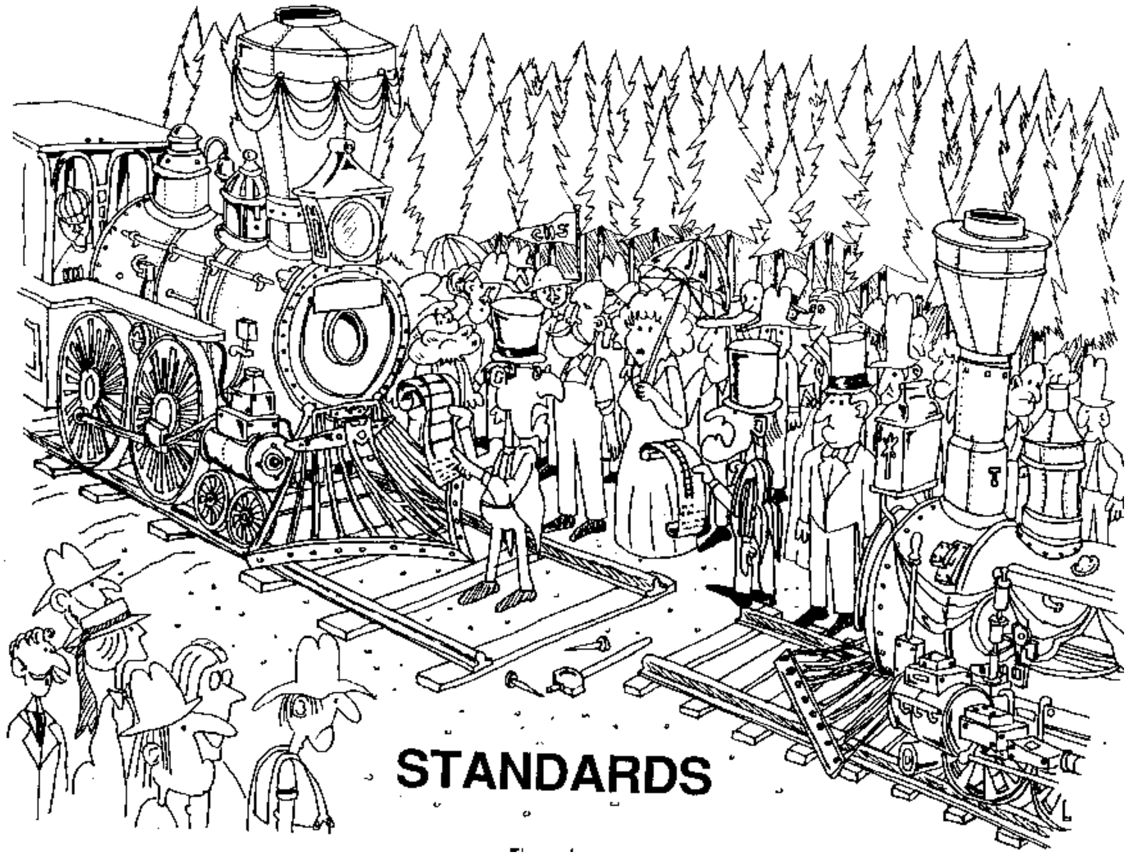
If it takes a decade to copy data to a new denser media, and if the life expectancy of the new media is also a decade, then there may not be enough time to transfer all the data to the new media.

Unless major breakthroughs occur in the speed of I/O transfers, or unless new scalable transfer processes are developed, it may not be possible to store permanently all the data that is collected in the future. Even if the process of migrating key information starts as soon as new storage media become available, the old storage media needs to have a life expectancy of at least twice the time needed to transfer data, in order to accommodate the explosive growth of our observational and computational systems.

5. Technology Shifts

Digital data often contains information that is meaningful solely to the software applications that created it. For example, most files created by Geographical Information Systems or word processing software embed proprietary format instructions describing content and structure. These files contain digital information that usually can only be interpreted by application software which operate in accordance with strict rules based upon highly specific definitions of the data content and structure. Unless these definitions are stored with the data in a known standard form, other applications that lack instructions for decoding them will be unable to use the information. Many commercial packages store information in complex proprietary architectures, which can only be interpreted by the software that created them. As a rule, descriptions of the file structures created by such applications are not readily accessible, resulting in data loss when the proprietary software is no longer available.

Brute-force decoding of these types of file systems is not an easy task, because the meaning of a file is not necessarily inherent in the bits that it contains. The contents of a file can sometimes be interpreted if it is self-contained,(i.e. if it includes encapsulated meta data and an adequate description of its architecture). Many information technologies embody such schemes, but regrettably, they often abandon them in the course of adopting new forms. This creates serious legacy problems by locking organizations into proprietary formats that may no longer be effective or versatile enough to meet evolving needs.



On a spring day in May 1869, the Central Pacific Railroad met the Union Pacific Railroad to span the American continent by rail. Fortunately, by this time the engineers had, by and large, settled upon a single track gauge standard. One might speculate on what might have happened if the 19th Century railroads had continued to pursue the policy of non-standardization, as is rampant today.

Some technology providers provide migration software to new formats and architectures; however, this may not be enough to solve the problem where large volumes of archived data are concerned. Many institutions are not able to incur the cost and effort of converting the legacy information to newer formats, and have no option but to **abandon** the older records. This information loss disrupts time series and trend measurements, which can result in wrong research or policy decisions made by analysts who have no choice but to work with inadequate information.

"Another half a million dollars and I'm sure we can make it go faster"

Another option is to maintain obsolete hardware and software to allow access to the legacy data. This adds to the cost of maintaining older holdings alongside the newer technologies. The retrofitting of obsolete hardware and software to the newer technologies has proven to be expensive, inefficient, time consuming and problematic. The alternative is to extract digital information in accordance with the specifications of the software that produced it. In principle, one does not have to operate this software if the rules of its operation can be described in a way that it does not depend on any particular computer system. With decoding rules and meta information encapsulated in the file itself, future generations of users should be able to recapture older digital information by recreating the behavior of the originating software.

6. Meta Data Information Loss

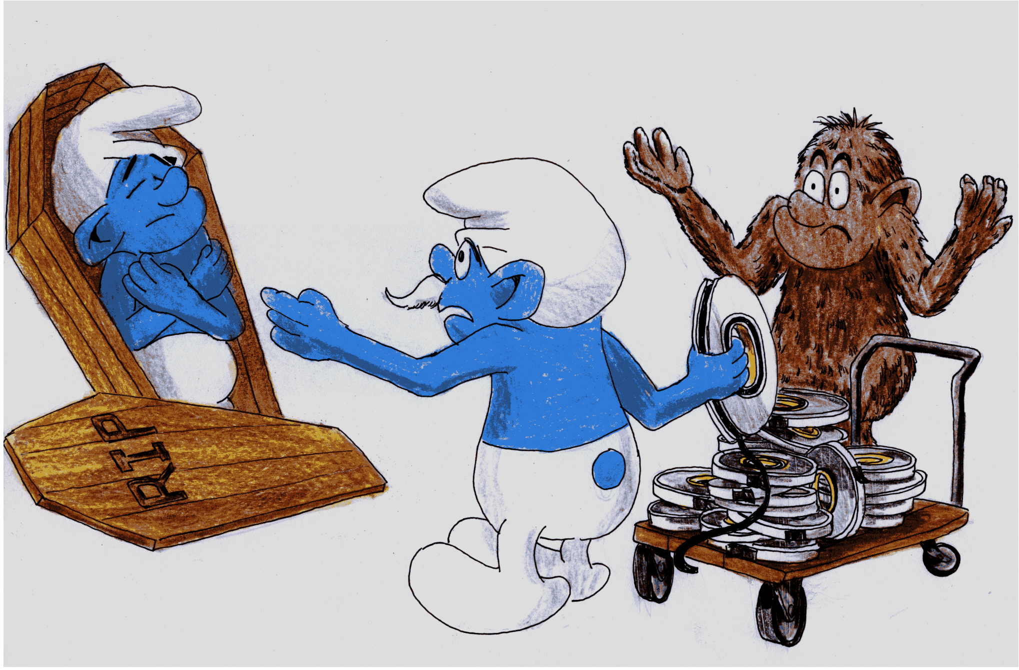
Meta information is the description of a particular data set: its provenance, its processing history, its format and other ancillary information required to realize the full value of the data. It should define file content and file structure, including details such as column names, column types, version dates and essential information about encoding schemes of the data such as coordinate systems, units, reference datums, and reference spheroids etc.

Proprietary GIS and image processing software often require that basic information about the data file be known, such as the spatial reference information. This information is usually not encapsulated in the file and is contained within the application software or ancillary file systems.

In some cases, the proprietary software does not have any spatial reference information at all and requires that the data be in some recognizable format that can be imported into the system. This is acceptable if only one dataset is being used but in many cases multiple disparate datasets are required for data analysis. The lack of appropriate metadata forces one to spend valuable time researching the lineage of the dataset. It is at this point that assumptions have to be made about the state of the data. Where did it come from? What specific geographic area does it represent? Is the data in its original form? What ellipsoid is it on? What datum is it on? The lack of this type information often leads to wrong conclusions determined from the data, based on wrong assumptions about the data.

Without proper meta information, assumptions have to be made with respect to the data, bringing into question its reliability and quality, which directly affects its interpretability. A basic question that has to be considered is the type of processing that has been applied to the data. Most application software do not have the capability to record or maintain metadata about digital geospatial data within the data file. This requires that the data manager or user record metadata information in separate files (or notes). This has the potential for meta data loss if this separate file or files should ever be compromised

Several standards for the collection and maintenance of geospatial metadata have already been developed (FGDC's Contents Standards for Geospatial Metadata). Although these standards are there, they are not always followed. The main problem lies in the inability to maintain this information within the datafile itself. Because of this problem, metadata can easily be misplaced or forgotten when transferring data from one organization or person to another.



“What do you mean he’s dead? He’s the only one who knows what this data is”

It has been argued that it is the responsibility of the data managers to maintain data and metadata, however this is not necessarily true with respect to the transfer of data. Several scenarios can occur where the data manager has no control over the data. For example, once the data is out of the hands of the data manager, he or she will likely have no idea of the sort of manipulations that have been performed on the data. This leads to the issue of metadata dependence on users. In many cases, the current user has to contact previous users of the data in order to acquire meta information that pertains to the historical context of the data. This has serious implications, if the previous user has retired, or worse, exited this plane of existence.

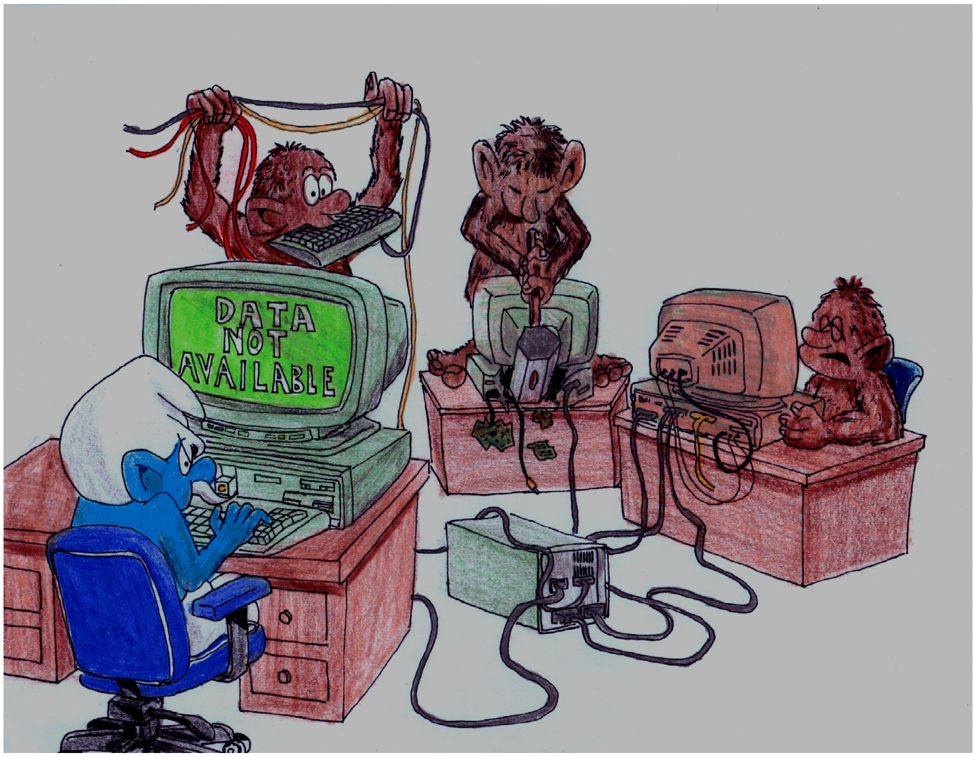
There is a series of pertinent information that must be stored with the data such as:

- What is the quality of data?
- What time frame was it collected in?
- What techniques and tools were used to obtain or manipulate it?
- Is it the original data or a generalization of multiple datasets?
- What is the scale resolution of the data?

If this type of metadata is not included with the dataset, it is necessary to locate the information. In some cases, the information may not be obtainable for various reasons, forcing assumptions to be made based on guesses. This reduces confidence in the quality of the data and affects its interpretation. Knowledge of the dataset’s pedigree gives the ability to provide a more accurate depiction to potential decision-makers.

7. Data Degradation Caused by Projection Conversions and Datum Adjustments : an example

Preliminary assessments of the outer limit of Canada's juridicial continental shelf have been based on the assembly and analysis of legacy data (MacNab et al, 1996). This information consisted of depth measurements that were collected by many agencies, with many platforms, on many different datums, using various technologies and various scale resolutions. In most cases, it was not clear whether the data had been normalized to a common datum or whether it had been left in its original form.This translated into considerable uncertainty about the depth and the positional accuracy throughout the data sets.



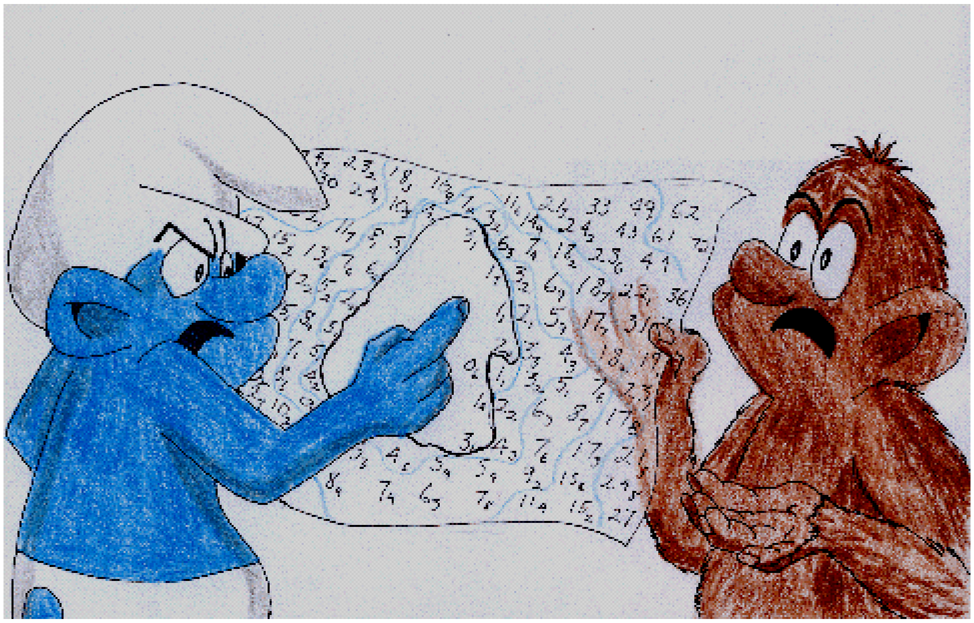
"Funny…. the data was there a minute ago"

There was no indication of the source of the data, let alone any information such as resolution, scale, ellipsoidal or datum information. The problem became obvious when the data sets held in the Canadian archives were merged and compared with data sets held elsewhere. Based on the shape and spacing of the points, duplicate tracks were identified; however, upon closer examination, they were observed to be slightly offset horizontally.

Another issue that became obvious during this exercise was that the points that could be identified as being the same had different depth values when obtained from different agencies. The probable cause was that there had been different sound velocity or vertical datum adjustments applied to the depth values. However, there was no meta information associated with the file indicating what or if velocity or tidal corrections had been applied to this data, leaving the data in a questionable state.

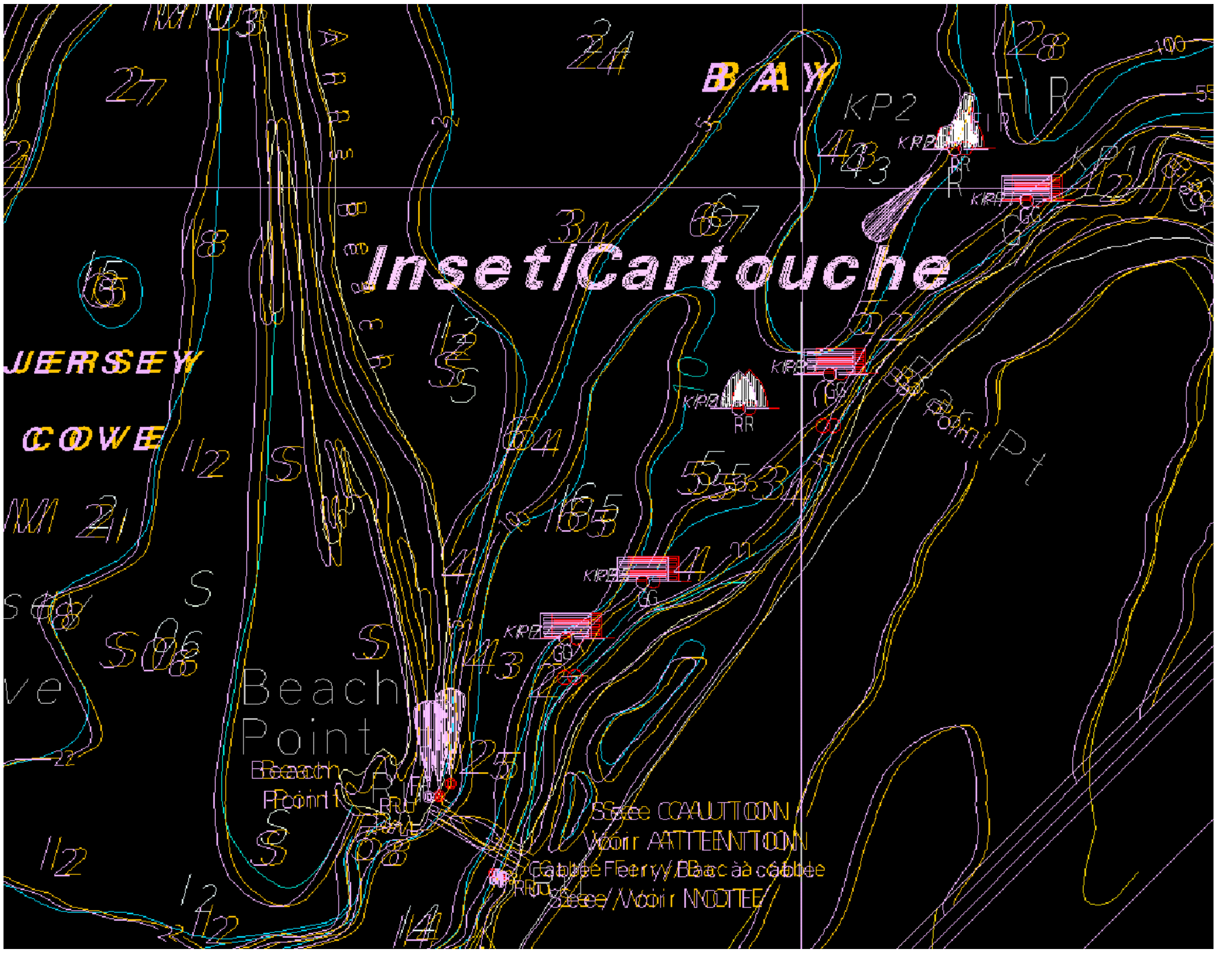
This leaves the dilemma of ascertaining which dataset is more valid. The lack of metadata in Canada's offshore dataset creates a huge problem. Should Canada ratify Article 76 – Law of the Sea, we will have to either use the data we currently have and risk losing some of the area that the country is entitled to due to nebulous information, or spend a great deal of money re-surveying Canada's offshore to verify the data sets.

A possible explanation for the variations in data could be differences in vertical and horizontal datums. A shift from NAD27 to NAD83 or WGS84, for instance can create discrepancies in the order of hundreds of meters.



"We've got to do something about depths plotting on land every time we change projections"

A second explanation is that conversions between projections (Mercator and UTM) can result in a loss of precision, caused by floating point roundoff errors in projection computations. This makes the case for storing spatial information in projectionless ellipsoidal coordinate rather than projected coordinates. Ellipsoidal coordinates can be depicted on any projection, which would virtually eliminate the degradation of spatial data caused by accumulation of roundoff errors through multiple conversions.

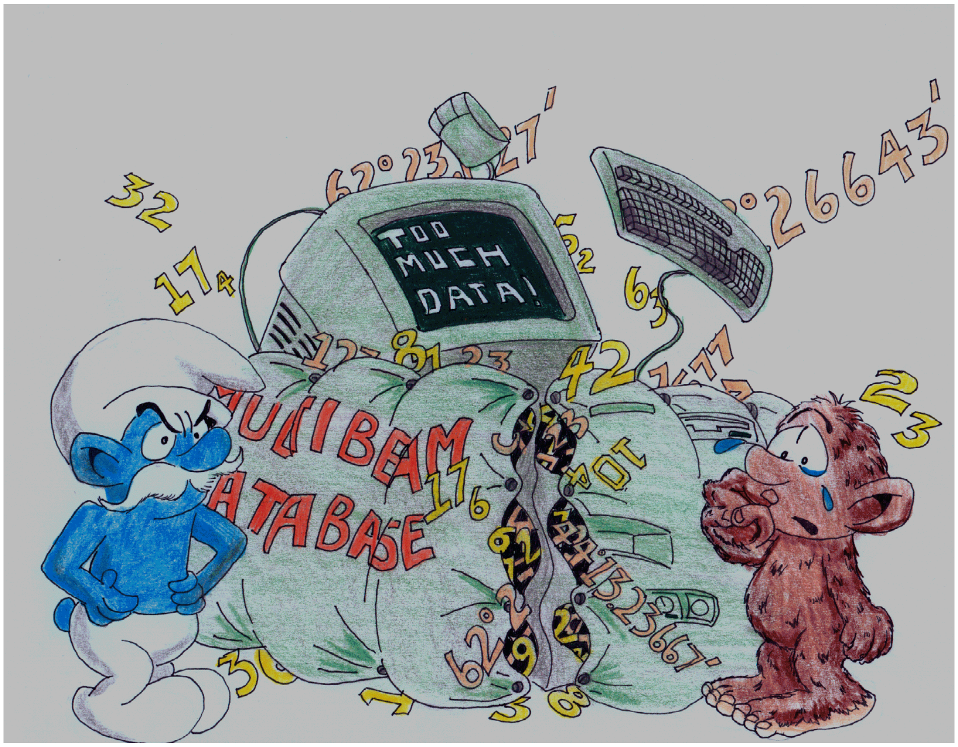


The above depiction is made up of three sets of data, generated from the same data points. They have been manipulated three times by the same GIS, causing severe round off problems as can be seen by the displacement of points and contour lines in the example. The initial collected positions no longer exist, only the transformed coordinates remain after each manipulation. The danger is that these degraded coordinates can be subsequently reentered into a spatial data warehouse without the user ever being aware that he has violated the data integrity of the warehouse.

This has severe implications if this level of degradation is applied to geodetic control, benchmarks etc. on which entire survey networks are based. If land parcels and territories are continuously shifted in an adhoc manner, one can imagine the legal implications of territorial rights, where land or sea areas, land parcels, national and provincial boundaries are insidiously being subjected to degradation in spatial databases that are legally administrated by national, international and municipal agencies.

8. Data Decimation and Scale

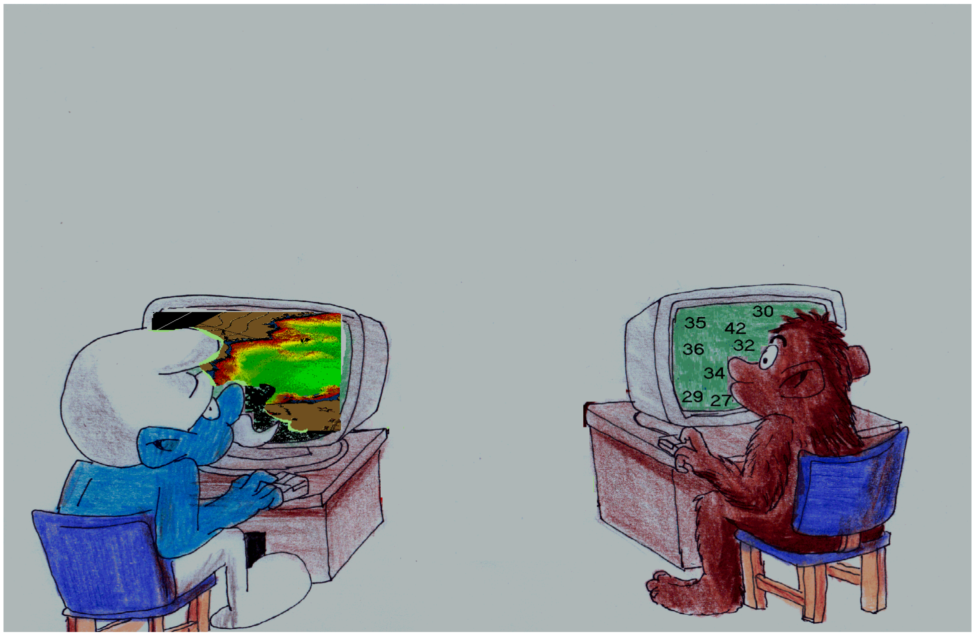
The use of digital geotemporal data plays a key role in the interpretation and analysis of environmental, geological and urban processes. Geotemporal data must contain locational, temporal and associated attributes necessary to formulate trends at the analysis level.



"Maybe we can make it fit by removing longitudes and depths"

Users of digital geotemporal data are often faced with a lack of essential attributes within their datasets. This is primarily due to the well intentioned but misguided attempts to reduce file size for data archival through attribute depletion or so called data decimation. Certain specific attributes are deemed unnecessary and are dropped or generalized in an effort to keep the file sizes manageable. An example of this was the lack of track identification in the legacy data sets used for the continental shelf delimitation. In order to compensate for this loss of information, a special purpose program had to be developed to infer artificial track numbers whenever there was a time or a distance gap. This was a best guess scenario, since the original track information was removed. In some cases, the original information was irrecoverable and the data was lost forever.

Retention of the original data set expands the use of the data, since it was already useful in the acquisition phase. It allows the data to have more functionality and makes it more usable for future purposes.



“How come when I zoom, only my numbers get bigger? ”

The issue of scale is highly important when it comes to data loss. In general, the relatively low density of data that is suitable for small scale maps is insufficient for detailed large scale maps. Conversely, the higher densities of data needed for detailed mapping of large scale maps are excessive for small scale maps. As a rule, detailed data can be generalized for the purposes of creating a large scale portrayal, once in this form it is impossible to return to the detailed data unless one actually maintains the original data. This locks the data to a generalized scale causing it to lose the functionality of scale independence.

Some have equated the word degraded to the word generalized as pertaining to data loss due to scale. To maintain a scale independent archive, it is necessary to store the data at the scale it was obtained or at the best scale obtainable. This allows the flexibility to generate maps or charts at various scales from source information.

9. Proprietary Data Formats

A significant issue that has become glaringly apparent is the wide variety of spatial data formats. The proprietary nature of the storage formats has lead to legal issues and ownership problems. Proprietary formats are, by their very nature, saleable commodities, subject to legal issues such as useage, royalties and transfer of ownership. The IP

(Intellectual Property) owners can sell or transfer rights on proprietary formats. They can refuse useage of their proprietary formats or demand exorbitant royalties for their use.

"Oh! No! He bought the rights to the proprietary format "

The danger is that foreign nationals could, in fact, legally control a nation's data holdings through IP ownership of proprietary formats and force policy decisions. Common solutions in the past were to convert data to common open ASCII formats so anyone could read or manipulate it as open architecture. However, as the data sets are becoming larger and more complex, ASCII is no longer an efficient way of managing or storing information. This has lead to a proliferation of unauthorized transformation software using proprietary formats, leading to problems with data integrity, data loss, data synchronization problems, as well as legal disputes over ownership. Many of these problems could be reduced through the standardization of open interoperable architectures for data storage.

10. Software Independent Storage for Long Term Archives Using the Self Defining Structure (SDS)

The Self-Defining Structure (SDS) is a specification that has been proposed to the ISOTC211 for encapsulated storage of data for long time archival of VERY, VERY large data sets.

The goals of SDS are to facilitate and ensure information survivability for long term archives by encapsulating standardized mechanisms within the file itself. This means the minimization of the dependence on applications and outside meta information for interpretation of data within the file.

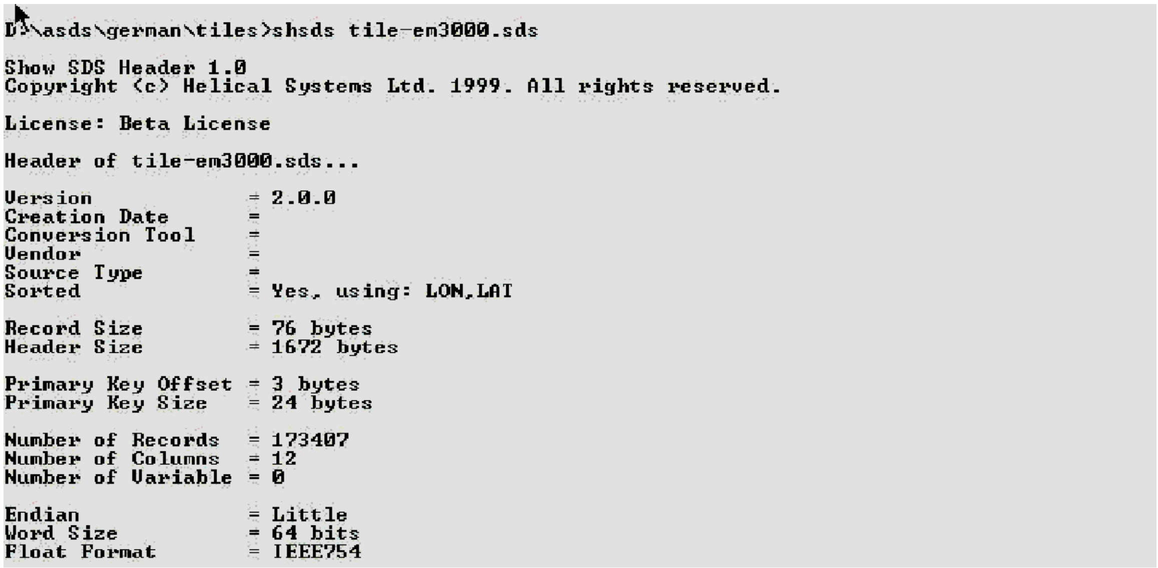
Some properties of the SDS are:

- Encapsulation of methods related to data elements in file.
- Encapsulation of meta information associated with data elements in file.
- Encapsulation of meta information about the file.
- Portability.
- Extendibility.
- Interoperability.
- Re-usability.
- Ease of Storage and Retrieval.
- Quick Access.
- Ease of maintenance.
- Longevity.
- Ease of organization and re-organization.

Object oriented concepts are used to provide the means of achieving these goals. The SDS architecture consists solely of objects. Every object by definition must have an associated class description specifying the format and methods to read, write, and manipulate it. In this manner, the SDS architecture inherits all the flexibility and power of the object oriented paradigm.

11. Features of the Self Defining Structure (SDS)

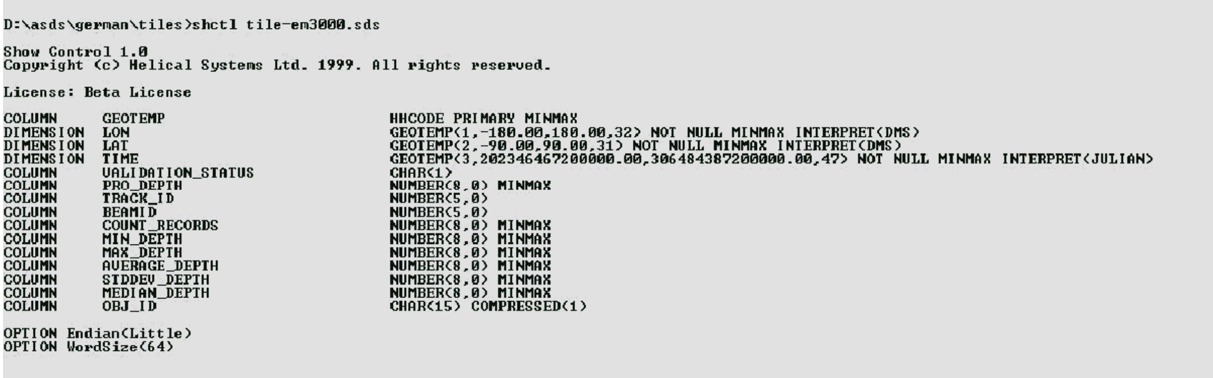
By design, SDS is not tied to the format or the content requirement of any particular application but features, instead, a high level of generality. The advantage of this concept is that the SDS supports data interoperability with a wide range of applications, permitting considerable latitude for the user in selecting the tools and processing techniques that are most appropriate to their needs.



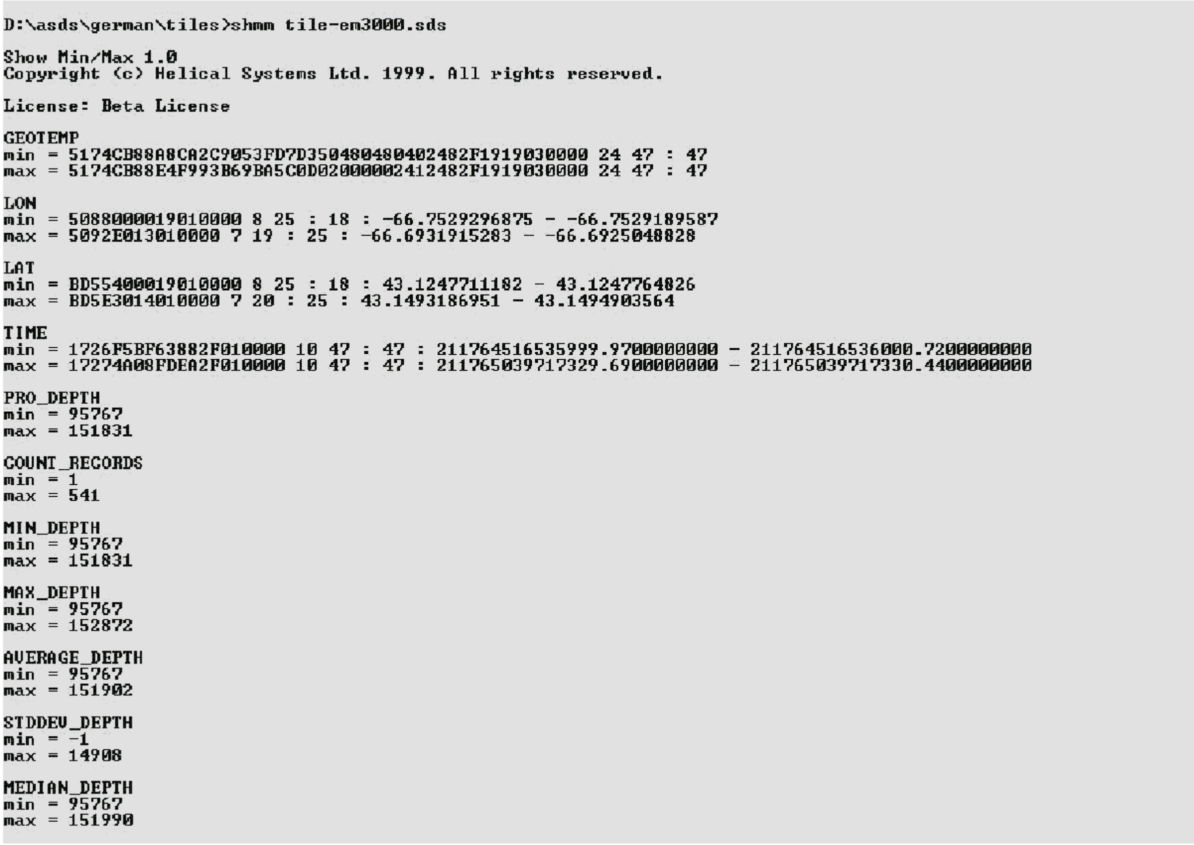
```
D:\asds\german\tiles>shsds tile-em3000.sds

Show SDS Header 1.0
Copyright (c) Helical Systems Ltd. 1999. All rights reserved.

License: Beta License

Header of tile-em3000.sds...

Version              = 2.0.0
Creation Date        =
Conversion Tool      =
Vendor               =
Source Type          =
Sorted               = Yes, using: LON,LAT

Record Size          = 76 bytes
Header Size          = 1672 bytes

Primary Key Offset   = 3 bytes
Primary Key Size     = 24 bytes

Number of Records    = 173407
Number of Columns    = 12
Number of Variable   = 0

Endian               = Little
Word Size            = 64 bits
Float Format         = IEEE754
```

The SDS architecture supports this by allowing the data to be fully described. For example, when defining a column, the method of interpretation may also be defined. This would describe to an application how to read and write the column information e.g. "The value in this column is a Gregorian calendar date in the format YYMMDD where 1900 YY is the year".

All definitions and characteristics of the data are encapsulated within the SDS file. The encapsulation of this information allows the definitions to survive within the file itself for future purposes. This allows a measure of data independence, allowing data interoperability between applications as well as encapsulated column information for long term archival.



```
D:\asds\german\tiles>shctl tile-em3000.sds

Show Control 1.0
Copyright (c) Helical Systems Ltd. 1999. All rights reserved.

License: Beta License

COLUMN      GEOTEMP                        HHCODE PRIMARY MINMAX
DIMENSION   LON                            GEOTEMP(1,-180.00,180.00,32) NOT NULL MINMAX INTERPRET(DMS)
DIMENSION   LAT                            GEOTEMP(2,-90.00,90.00,31) NOT NULL MINMAX INTERPRET(DMS)
DIMENSION   TIME                           GEOTEMP(3,202346467200000.00,306484387200000.00,47) NOT NULL MINMAX INTERPRET(JULIAN)
COLUMN      VALIDATION_STATUS              CHAR(1)
COLUMN      PRO_DEPTH                      NUMBER(8,0) MINMAX
COLUMN      TRACK_ID                       NUMBER(5,0)
COLUMN      BEAMID                         NUMBER(5,0)
COLUMN      COUNT_RECORDS                  NUMBER(8,0) MINMAX
COLUMN      MIN_DEPTH                      NUMBER(8,0) MINMAX
COLUMN      MAX_DEPTH                      NUMBER(8,0) MINMAX
COLUMN      AVERAGE_DEPTH                  NUMBER(8,0) MINMAX
COLUMN      STDDEV_DEPTH                   NUMBER(8,0) MINMAX
COLUMN      MEDIAN_DEPTH                   NUMBER(8,0) MINMAX
COLUMN      OBJ_ID                         CHAR(15) COMPRESSED(1)

OPTION Endian(Little)
OPTION WordSize(64)
```



```
D:\asds\german\tiles>shmm tile-em3000.sds

Show Min/Max 1.0
Copyright (c) Helical Systems Ltd. 1999. All rights reserved.

License: Beta License

GEOTEMP
min = 5174CB88A8CA2C9053FD7D350480480402482F1919030000 24 47 : 47
max = 5174CB88E4F993B69BA5C0D020000002412482F1919030000 24 47 : 47

LON
min = 5088000019010000 8 25 : 18 : -66.7529296875 - -66.7529189587
max = 5092E013010000 7 19 : 25 : -66.6931915283 - -66.6925048828

LAT
min = BD55400019010000 8 25 : 18 : 43.1247711182 - 43.1247764826
max = BD5E3014010000 7 20 : 25 : 43.1493186951 - 43.1494903564

TIME
min = 1726F5BF63882F010000 10 47 : 47 : 211764516535999.9700000000 - 211764516536000.7200000000
max = 17274A08FDEA2F010000 10 47 : 47 : 211765039717329.6900000000 - 211765039717330.4400000000

PRO_DEPTH
min = 95767
max = 151831

COUNT_RECORDS
min = 1
max = 541

MIN_DEPTH
min = 95767
max = 151831

MAX_DEPTH
min = 95767
max = 152872

AVERAGE_DEPTH
min = 95767
max = 151902

STDDEV_DEPTH
min = -1
max = 14908

MEDIAN_DEPTH
min = 95767
max = 151990
```

The SDS architecture supports the calculation of minimum and maximum values for designated columns within the SDS file. By not having to perform intensive full file scans, applications are provided a very fast filter capability for any query based on spatial, temporal, or attribute based constraints.

Repeated data stored in a SDS file may be automatically normalized resulting in an overall reduction of the file size. The SDS architecture allows column values to be represented as byte values referenced to a dictionary. This technique allows the file size to remain manageable without compromising attribution through data decimation.

The SDS architecture also addresses the idea of smart raster (Varma 98) as it provides the ability to store tessellations with associated attributes and RGB colour values within the same file.

The final SDS design will encapsulate current data models using object oriented concepts. It should emulate the concept of a book, where all meta information relevant to the content and context of the file are resident within the same framework. Not only does this provide better interoperability between applications in the immediate time frame, but it also prevents information loss in the future.

12. Distributing the Archive over several machines improves Data transfer rates.

Several companies are currently exploring options to alleviate the media to media transfer bottleneck problem by partitioning and distributing data archives over several machines using fast network technology and synchronized data dictionaries. The archive managed over several machines can be treated as a single data archive for the purposes of data storage and data extraction. However, each individual machine can also provide secular throughput and maintenance of independent file transfer or file storage without impacting the overall archive. This allows the simultaneous transfer of data to newer media on a machine by machine basis in parallel.

The distributed archive architecture not only has the potential of solving the media to media transfer problem but, in fact, lowers the cost of the archive by allowing the use of several cheaper mid- range machines in parallel, as opposed to expensive million dollar servers.

Oh no! I'm five dollars short of a million for the new server

The distributed archive capability also reduces the cost of the maintainance of component hardware due to technology improvements. The organization can cost effectively add or replace low cost individual component machines on a needs basis without impacting the overall archive. The replacement of expensive million dollar servers every few years due to obsolescence can be quite a traumatic experience in this already resource strapped world.

14. Conclusion

The key objective of creating knowledge-based information systems is to ensure the long-term survivability of digital data within large sized archives.
It is critical to ensure the survivability, reliability and evolvability of these systems before a non-recoverable data loss occurs. This is the information that will be the basis for future data analysis, research and policy decisions.
The essence of the crisis can be stated as follows: Even if new technology and parallel archive capability increase the media to media data transfer rates to a level that is comparable with the improvements in data storage technologies, the life expectancy of the information in the archive itself is in jeopardy due to the constant change of data architectures dictated by proprietary software applications.

There is an urgent need to address the design and implementation of open stable file architectures, such as SDS, under the guidance of international organizations, before the volume of data itself becomes an impediment to the survivability of digital information into the new millenium. .

Acknowledgements:

Thanks to Mike Collins for the cartoon Illustrations
The creator of the Smurfs is Pierre Culliford or Peyo

Bibliography:

Rothenberg J. Ensuring the longevity of Digital Documentation
Scientific American 1995

Halem M. et al Technology Assessment of High Capacity Data Storage Systems:
Can we avoid a Data Survivability Crisis. Feb 1999
White Paper Earth and Space Data Computing Division.

Varma H. et al Object Linked (Smart) Raster June 1999 **ISO/TC 211/WG1**
Image and Gridded Data on Geomatics

MacNab R. et al An Improved Bathymetric Portrayal of the Northwest Atlantic for Use in Delimiting
the Juridicial Continental Shelf According to Arcticle 76 of the Law of the Sea
Proceedings of the Canadian Hydrographic Conference Halifax June 3-5 1996
pp 8-13