GEOPHYSICAL DATA STEWARDSHIP IN THE 21ST CENTURY AT THE NATIONAL GEOPHYSICAL DATA CENTER (NGDC)

E A Kihn[*] and C G Fox

NOAA/ National Geophysical Data Center, 325 Broadway E/GC, Boulder CO, USA
**Email:* Eric.A.Kihn@noaa.gov

ABSTRACT

The World Data Center for Geophysics in Boulder, Colorado is hosted by the National Geophysical Data Center (NGDC). NGDC's vision is to be the world's leading provider of geophysical and environmental data, information, and products. NGDC's mission is to provide long-term scientific data stewardship for geophysical data, ensuring quality, integrity, and accessibility. Faced with ever expanding data volumes and types of data, NGDC is developing more innovative techniques for science data stewardship based in part on data mining and fuzzy logic. Use of these techniques will allow NGDC to more effectively provide data stewardship for its own scientific data archives and perhaps the broader World Data System.

Keywords: Geophysics, Marine geology, Space weather

1 INTRODUCTION

The National Geophysical Data Center (NGDC) (http://www.ngdc.noaa.gov/ngdcinfo/aboutngdc.html) is one of three data centers operated by The National Atmospheric and Oceanic Administration (NOAA) to archive and disseminate data collected in executing its environmental mission. NGDC has two primary science divisions each focused on a different domain. The Solar and Terrestrial Physics (STP) division, which focuses on space related and space derived products and information, and the Marine Geology and Geophysics (MGG) Division, which focuses primarily on data from the sea floor as well as main field magnetics. A sample listing of the data and applications from each is available in Table 1.

Table 1. Sample data products and their application areas

DATA TYPES	APPLICATIONS
Bathymetry	Natural Hazards Assessment
Digital Elevation Models	& Economic Impact
Gravity & Magnetics	Tsunami Inundation Modeling
Ocean Drilling	Ocean Mapping
Seismic Reflection	Defense Applications
Sea Floor Composition	Cable & Pipeline Routing
Bottom Pressure Recorder (BPR) Data	Minerals Exploration
Natural Hazards Photos	Fisheries; HabitatsGlobal Change
Significant Earthquakes	Research
Volcanic Deposits	Climate & Global Change
	Satellite Operations
Solar Imagery	Space Weather Models
NOAA/TIROS Particles	Electrical Power Networks
GOES Particles and Fields	Radio Communications
Spacecraft Anomalies	Education
Geomagnetic Variations	Remote Sensing
Auroral Images	Global Positioning Satellites
Ionospheric Parameters	Solar Research
DMSP Particles and Fields	Space Research
Solar Radiation	

A primary component of NGDC's mission is to provide scientific stewardship for the data archived at the center. Here "scientific stewardship" means that in addition to preserving the data for the long term, NGDC focuses on providing calibrated data sets that can reach a broader audience, creating products from raw data and thereby exposing the data to a larger audience; providing long term quality control for data sets to create "research quality holdings"; and finally propogating the knowledge derived from the data to the community at large. As can be seen in Figure 1, because each of the higher level activities is labor intensive, it is performed on a proportionally smaller percentage of the overall data archive, thereby reducing the return on investment made in archiving the data. NGDC is developing tools and techniques that allow the center to address more of the data at a higher level without increasing overall staff even in the face of increasing data volumes and diversity. The goal is to develop automated "expert systems" that provide stewardship functions without the need for direct staff involvement. The sections below describe the NGDC vision and some early implementations in pursuit of more automated and improved data stewardship.

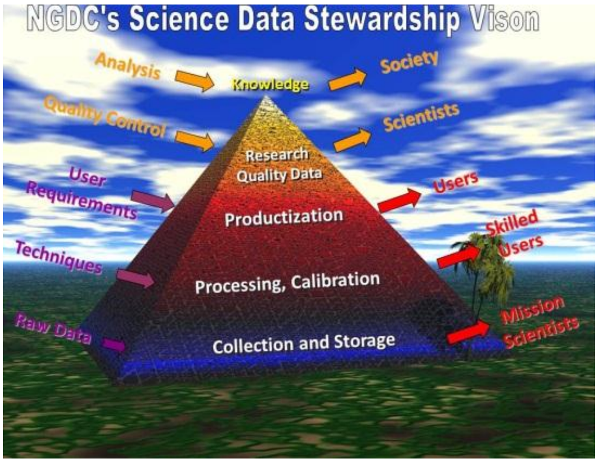

Figure 1. The stewardship pyramid showing decreasing volume for higher order products

2 DATA MINING

Data mining is one possible solution in support of stewardship activites. By data mining we mean using mathematical and computational tools to extract previously unknown, and potentially useful, information from the archived data. Data mining uses techniques such as machine learning and statistical analysis to summarize and present knowledge in a form that is easily comprehensible to humans. By filtering through the vast achives and pointing trained scientists to the more interesting bits of information, data mining enables management of larger and more diverse archives. Some possible applications of these techniques are summarized in Table 2. The first two are addressed with specific examples below.

Table 2. Applications of data mining

Applications of Data Mining
• **Data quality control**
• **Human linguistic translation**
• Event and trend detection
• Data classification
• Forecast
• Deviation detection

3 HUMAN LINGUISTIC TRANSLATION

When attempting to mine data for information, we find that natural language is not easily translated into the more computer-friendly terms of simply 0s and 1s. However, natural language is typically how scientists prefer to ask questions when interacting with data: Is the sample "hotter" on average?; Is this observation outside of the "norm"?; Is the sample "changing" with time? Fuzzy logic lets us map human thought and language into computer functions much closer to the way the brain works. We can aggregate data and form a number of partial truths, which we consider when certain thresholds are exceeded, initiating an action such as flagging the data as suspect or identifying a significant trend. Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth -- truth values between "completely true" and "completely false". It was introduced by Dr. Lotfi Zadeh (Zadeh, 1965) of UC/Berkeley in the 1960s as a means to model the uncertainty of natural language The use of "fuzzy" logic allows automated systems to capture some of the natural thought process of a data manager and to apply it to an archive. Applying these techniques, one can search an entire 40-year archive for events described by "high" winds, "average" temperature, and "about" 60% humidty (perhaps a storm description) and quickly identify when such events are occuring, detect any changes over time, and display the results to a user (Figure 2). Notice that because the language used is natural, the same query would work for data in Alaska or Florida although what constitutes "average" temperature is obviously quite different between the two. Natural language processing is key to handling large and diverse data volumes and will be expanded at NGDC as ever more automated systems are fielded.

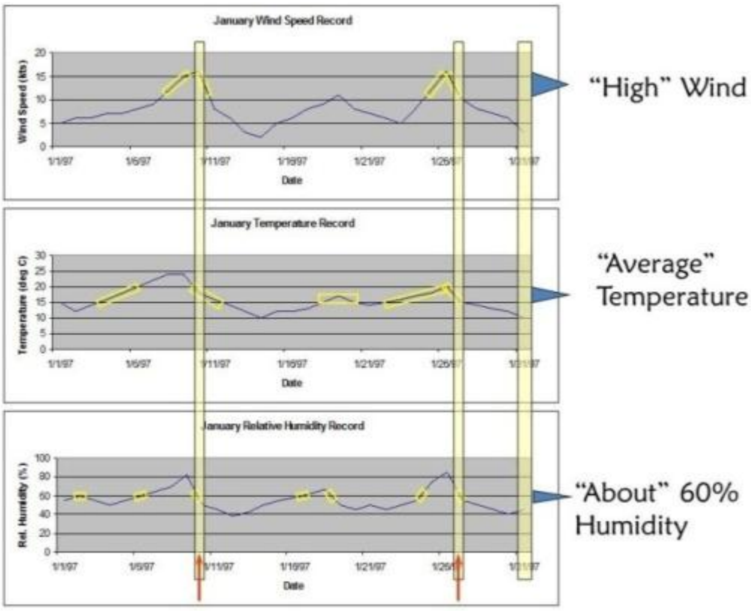

Figure 2. A sample search for a typical weather event

4 DATA QUALITY CONTROL

The Space Weather Reanalysis (SWR) (Kihn, 2007) is a long term reanalysis of space weather data that requires careful quality control of a huge volume of diverse data. The SWR involves taking raw observational data and processing it through linked physical models that produce a higher order product capable of summarizing the state of the space environment. A single instance of bad data can have ripple effects throughout the entire model run. Working with satellite and station data in particular can be tricky, with spikes, baseline shifts, and dropouts all prominent in the data stream (Figure 3). In a typical small scale study it would be possible for a researcher to hand screen the data, but here the data volume requires the application of "intelligent" computer techniques, based on fuzzy-logic, neural computing, and other mathematical functions. In particular for this application of data quality control, NGDC developed a system capable of "peer matching"; that is, each station was analyzed to determine a group of peer stations based on location, instrument type, and dynamic range. The data mining application was then set to look at the entire 15 year data stream for instances when a given station observed data "unlike" its peers. This much smaller subset of data could then be reviewed directly by an analyst. Notice that in this instance the data mining helps in two ways: by determining a set of peer stations and by allowing a linguistic search for data "unlike" its peers. Application of these techniques allowed for integration of over 15 years of data into the model runs but also left behind a vastly improved data archive, with each station and observation having been screened for quality. NGDC will look to expand usage of such systems as data volumes and diversity increase.

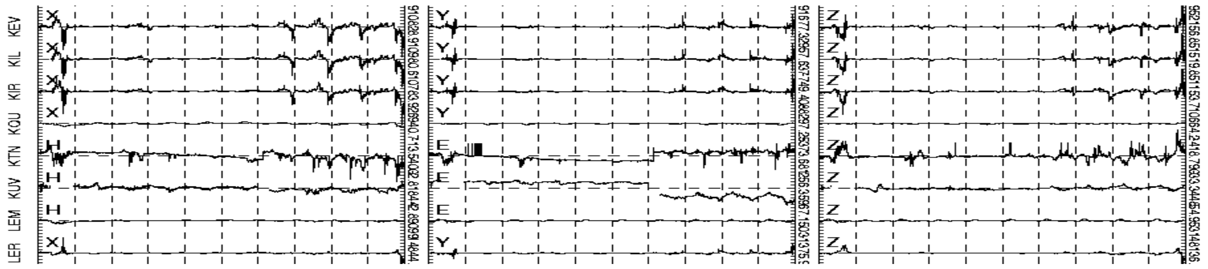

Figure 3. Sample magnetometer data used by the SWR project

5 CONCLUSION

It is clear that increasing data volumes and data diversity demand new tools and methods. While the amount of data and number of data sets tends to increase exponentially (Figure 4), the number of staff available to manage the data remains level. Mathematical methods exist that provide analysis, classification, and forecast methods for large data volumes, specifically the data mining and fuzzy logic systems mentioned above.. In particular, fuzzy-based systems hold great promise as knowledge extraction tools, allowing for better information extraction from the vast and diverse archives available. One of the greatest challenges facing science in the coming years is how to effectively utilize the data archived not only at a single center but available across a distributed network such as the World Data System. The techniques described above can play an important role in this effort by better integrating the data and helping scientists to focus on the most relevant bits. Without the development of such tools and systems, the extensive data archives of the World Data System will be vastly underutilized and their scientific potential squandered.

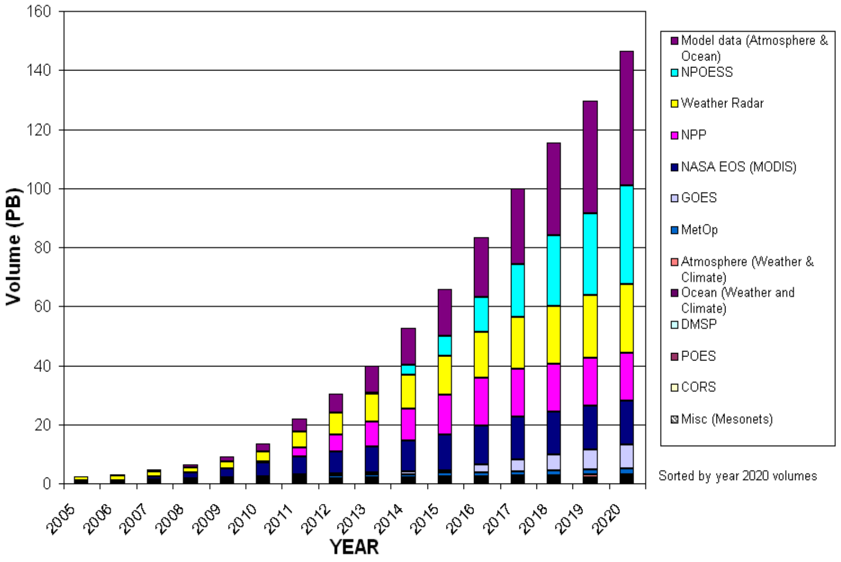

Figure 4. Expected data volume growth at NGDC

6 REFERENCES

Kihn, E.A., Ridley A.J., & Zhizhin, M. (2007) The Space Weather Reanalysis. *Materials of the International Conference '50th Anniversary of the International Geophysical Year and Electronic Geophysical Year'*, GC RAS, Moscow, doi:10.2205/2007-IGY50conf.

Zadeh, L.A. (1965) *Information and Control 8* (3), pp 338–353.

(Article history: Available online 10 April 2013)