TOWARD INTELLIGENT AND SUSTAINABLE ARCHIVING OF GEO-SPATIAL DATA

Gilbert L. Rochon[1], Dev Niyogi[2], Alok Chaturvedi[3], U.C. Mohanty[4] , Larry Biehl[5], Joseph E. Quansah[6], Krishna Madhavan[7], Rajarathinam Arangarasan[8], Maruti Ram Ponaganti[9], Hussam Nour[10], Aisha Chini Reed[11], Anil Kumar[12] and Hsin-I Chang[13]

[1]Purdue University, Chief Scientist, Rosen Center for Advanced Computing, Associate Vice President for IT Collaborative Research and Director, Purdue Terrestrial Observatory rochon@purdue.edu; [2]Purdue University, Assistant Professor of Agronomy and Earth & Atmospheric Sciences and Indiana State Climatologist, West Lafayette, Indiana USA dniyogi@purdue.edu; [3]Purdue University, Professor, Krannert School of Management & Director, Purdue Homeland Security Institute; [4]Professor, Center for Atmospheric Sciences, Indian Institute of Technology, New Delhi, India; [5]Purdue Terrestrial Observatory, Systems Manager; [6]Purdue University, Doctoral Student, Dept. of Agricultural & Biological Engineering and Graduate Research Assistant, Purdue Terrestrial Observatory; [7]Senior Scientist, Rosen Center for Advanced Computing, Information Technology at Purdue; [8]Purdue University, Visualization and Computer Graphics Applications Engineer, Assistant Professor of Computer Graphics Technology; [9]Purdue University; Graduate Research Assistant-IndianaView, School of Mechanical Engineering; [10]Purdue University, Graduate Research Assistant, Krannert School of Management; [11]Purdue University, Graduate Student, Department of Earth and Atmospheric Sciences and Exec. VP & Treasurer Intern; [12]Purdue University, Postdoctoral Researcher, Department of Earth and Atmospheric Sciences, and National Center for Atmospheric Research (NCAR); [13]Purdue University, Graduate Research Assistant, Department of Earth & Atmospheric Sciences

ISPRS International Symposium on Geospatial Databases for Sustainable Development, Sep. 27-30, 2006, Goa, India

KEY WORDS: Heterogeneous Database, Environmental Systems, Remote Sensing, Visualization, Education, Purdue University.

ABSTRACT :

A critical assessment of the data archive system at Purdue University in West Lafayette, Indiana, USA, is detailed with specific reference to those aspects that may be appropriate within a sustainable development context. The Purdue Terrestrial Observatory (PTO) benefits from an archive of satellite data, amassed over the past forty years by the Laboratory for Applications of Remote Sensing (LARS), real-time multi-sensor satellite groundstations, a recent five petabyte upgrade to the storage system, an array of available Linux clusters and supercomputers for rapid data analysis, fusion and mining, a state-of-the-science data Visualization and Perceptualization Center, real-time nation-wide access to NEXRAD/WSR-88D Doppler Radar installations, an elaborate network of high bandwidth connectivity to other academic research centers and federal government laboratories throughout the United States, high-resolution weather forecast output from a variety of environmental models, international partnerships to facilitate data sharing and, of singular importance, an affiliated interdisciplinary cadre of thirty-five scientists from twenty academic departments, supported by a fiscally committed administration, that believes in the importance of the PTO attaining preeminence in its field. Notwithstanding all these atypical benefits, the PTO faces some of the same technological and political constraints and challenges faced by most geospatial archives. The authors examine the state-of-the-science for geo-spatial archiving longevity and security and offer recommendations with respect to a strategy for incorporating expert system enhanced or "intelligent" archiving, while maintaining fault-tolerance and both scalability and the level of rapid access needed for time-sensitive early warning and disaster mitigation.

In order to accurately "hindcast," "nowcast" or "forecast," one requires the benefit of a long-term consistent archive. *In situ* oceanic measurements, geological records, limnological cores, polar ice cores, tree ring analysis, *inter alia*, all provide a basis for long-term assessments; however, the media on which data is stored, including handwritten documents on acidic paper as well as BIL and BSQ tapes, is largely ephemeral and in many instances urgently requires restoration, preservation of antique tape readers and digitization to more permanent media. The satellite data archive suffers from having been of short duration. Even including the de-classified military spy satellite data (e.g. Corona, Lanyard and Argon), we can only go back to 1960, and then without available global coverage, or uniformity of spatial or spectral resolution, georeferencing or regularity in re-visit time. The political decision not to safeguard full continuity for long-standing operational systems, such as Landsat MSS & TM places the scientific community at an extreme disadvantage.

INTRODUCTION

LARS Image Data Archive at Purdue University

The original aircraft and satellite (Landsat MSS & TM) remote sensing image data archive at Purdue University's Laboratory for Applications of Remote Sensing (LARS) consisted of a few thousand images stored on several thousand 9-track tapes. When funding for agricultural remote sensing applications was reduced in the early 1980's, LARS moved from the offices in Purdue's Research Park to campus. Some of these image date tapes were sent to researchers at other universities within Indiana. Many were moved to long term storage in Purdue Library system. However, in time, these tapes were disposed of as the space was needed by the library for other items. The image data on the tapes that were from areas over Indiana were copied to optical disks during the 1990's before the 9-track magnetic tapes were disposed of. A catalog of these images were made accessible via the web at the LARS web site (http://www.lars.purdue.edu/home/image_data/image_data.htm). These images are stored on around 200 cd's along with aircraft and other satellite remote sensing image collected during the 1990's thru the current time for research activities. During the late 80's and 90's there was no directly funded project or task to archive and keep track of remote sensing image data at Purdue. What was done was accomplished was between the "cracks".

More recently since 2004, the Landsat MSS and Thematic Mapper images stored on these CD's have now been made available "on-line" using the Tera- Grid project and a portal to the images that are part of IndianaView (http://www.indianaview.org/) which is a member of AmericaView Inc. (http://www.americaview.org/). AmericaView is a nationwide program that focuses on promoting the use of satellite remote sensing data and technologies in support of applied research, K-16 education, workforce development, and technology transfer. The Landsat data in the archives at Purdue University have been combined with that stored at Indiana University and Indiana University - Purdue University at Indianapolis. The archive is currently accessible via the IndianaView portal and consists of over 130 Landsat Thematic Mapper scenes for Indiana.

Purdue Terrestrial Observatory

Very recently (2005 and 2006), two satellite receiving stations have been installed at Purdue University as part of the Purdue Terrestrial Observatory (http://www.itap.purdue.edu/pto/). One is a geostationary antenna that receives data from GOES-12, also called GOES-East. The other is a 4.5 meter tracking antenna within a 5.5 meter radome from SeaSpace. This antenna receives image data from the Terra and Aqua MODIS sensors, the AVHRR sensors on several NOAA satellites and the MISVR sensor on the Feng-Yun-1d satellite. The tracking antenna has also received data from IRS' P4 (Oceansat) satellite. These receiving stations present a whole new scale and set of decisions for archiving remote sensing image data. What should be archived for real-time research? For how long should the images be stored? What should be kept on spinning disk? What can be kept on tape storage?

The amount of raw MODIS data received per day for archiving is 9 to 10 GB. These are the data sets that can be pulled back up when needed to do create specific data products. The amount of raw AVHRR data received per day for archiving is 1.0 GB and MVISR data is 0.5 GB.

ARCHIVAL GEO-SPATIAL DATA SEARCH AND ORDERING

The ultimate benefit of a highly intelligent and sustainable archiving of geo-spatial data is the ease of availability of geo-spatial data to users of all levels of expertise in spatial analysis. It is expected that both professionals and non professional users of such spatial data set can have quick and easy assess to data set of interest.

In the US alone, there over 250 spatial data servers, providing geospatial data through clearinghouses with links to many other international clearinghouses and many other research affiliated geospatial data archive systems worldwide. These systems provide all forms of spatial data for Geographic Information Systems (GIS) and image processing analysis via various mediums such as online FTP. Some common search portals and entry points to these geospatial data clearinghouse in the US are i) ESRI, ii) FGDC-National Spatial Data Infrastructure (NSDI), iii) NOAA Coastal Service Center (CSC), iv) Natural Resources Conservation Service (NRCS), v) Alaska Geographic Data Committee (AGDC), vi) Earth Resource Observation and Science (EROS) Data Center (EDC) (FGDC,2006).

CHALLENGES TO DATA SEARCH

While there are various sources of reservoir of geo-spatial data, there are a number of challenges one has to go through in order to have access to the right data set at the lowest possible cost. The first and most important issue is identifying the home pages and search portals for various geospatial data and satellite products. The easiest approach is starting with government geospatial institutions such as NASA, USGS, NOAA, USDA, National Geospatial Data Clearinghouse and universities with geospatial archival systems. These starting points lead to other interfaces where spatial data can be searched based on description such as location, granule numbers, specific time period, data type of interest, and resolution.

However searching and finding data is not as easy as it sounds since it can sometimes take a very long time (weeks and months) to obtain the correct source and perfect dataset for the area of interest. Some of these web interfaces are very complex and requires the user to read and understand complicated download instructions from manuals. Getting the right granule number for one’s specific area of interest is also a difficult task. Most archival and data distribution system do not provide a key on the granule orientation as used in the satellite scanning process. In most cases, a query for a specified research area produce so much data granules, leaving the user to painfully identify data specific to his area of interest. The issue of data availability is also of concern as a time consuming search could eventually lead to no useful data for the area of concern.

Apart from the complexity of most search portals, there is no immediate customer help for most archival and data distribution systems. Response to customer enquiries could take from hours to weeks in extreme cases. Another important setback to the utilization of geo-spatial data from most archives is the cost of obtaining these data. The cost could be free for coarse resolution data such as MODIS to very expensive high resolution geospatial data such as QuickBird, and IKONOS.

The last challenge to geospatial data users is suitable software packages for processing and manipulating the different datasets. Most of the geo-spatial datasets require specific software packages to be able to obtain maximum processing benefits from the datasets. The MODIS HDF-EOS data format for instance, comes with so many specialized software tools and algorithms for visualization, reprojection and other analysis. Such software tools for MODIS data alone include MOVAS, MSPHINX, HDFLOOK, geoview, simap, MODIS Reprojection Tool (MRT) and others (NASA, 2006). While there are standard software like ENVI, ERDAS IMAGINE, Idrisi etc, there are also so many required software tools for different geospatial datasets, making it tedious in identifying the appropriate software for each dataset. The need for users to learn and use these added software in addition to the standard processing software could be intimidating to non professional geo-spatial users.

IMPROVING EASE OF DATA SEARCH

A number of things could be done to make data search from archival systems more user-friendly and less difficult. Most web interfaces need to be simplified and customer assistance programs improved. Regular training conferences for users will help improve customer knowledge and ease of data access. The development of so many geo-spatial data archiving systems also means there could be reduction in cost of geospatial data. At present the cost of geo-spatial data is too expensive and this prevents many data users from seeking the appropriate data for their application.

EDUCATION, TRAINING AND SIMULATION The emergence of integrated approaches to environmental management and applications of remote sensing datasets for heterogeneous databases has exposed the critical need for new and innovative strategies for training the next generation of scientists who can utilize the multisensor data for science, analysis, and decision making. One of the most critical characteristics that need to be accounted for is that any education and training in this area needs to resemble and mimic closely the real world. Authenticity of the learning experience that allows learners to solve complex problems and use hands-on techniques will be valuable in ensuring transfer of skills from the educational setting to real world contexts. Madhavan et al. (2006) point out the need for utilizing 3D simulations and visualizations as an intrinsic part of the training process for handling bio-terror crises scenarios. The field of disaster management and mitigation, for instance, can be considered the larger problem space to which this approach of using 3D simulations and immersive environments can be used. However, the availability of such advanced and expensive environments in many parts of the world is limited and we have to look at lower cost ways for delivering similar training and educational opportunities.

The use of simulations has been long considered a successful strategy for incorporating near real-world scenarios into the educational curricula (Vieth et al., 1998; Reigeluth and Schwartz, 1989; Mackenzie et al., 2001). While the high-end of simulation use for training requires the availability of complex hardware tools, there are efforts such as the U.S. National Science Foundation funded Network for Computational Nanotechnology (NCN). The front-facing side of the NCN is a science gateway called the nanoHUB (http://www.nanohub.org), which provides users with easy access to advanced, industrial strength simulation tools that also tie into the national cyberinfrastructure fabric. The use of such science gateways for disaster management and mitigation could be invaluable to not only learners, but also to scientists. The emergence of science gateways to deliver advanced simulation for research and education has also highlighted the need for communities to have a single point of computation, data, and tool access. The ability to collaborate and contribute community approved and reviewed code to bridge education and research will be a key ingredient that will determine the success of disaster management and mitigation efforts.

In the Indian context, the technology infrastructure to deliver training that utilizes tremendous amounts of data and advanced visualization tools that rely on heavy computational environments is still developing. The use of science gateways as an aggregation point for the various data and analytic tools is not only feasible, but is also a practical, viable decision at this time point. Such efforts not only lower entry barriers to the field, but also provide educators with continuous engagement to the science and the scientific community. Similar efforts to train educators in the computational sciences in the United States deployed as part of the IEEE Supercomputing conference's education program – require educators to engage with materials related to the application of as part of a year-long program. The program structure allows participants to use web portals that utilize the open source course management Sakai to learn, practice various cutting-edge techniques, and for assessment. Participants also receive significant face-to-face with expert instructors. The core of all such efforts is the use of science gateways that allow research materials to be disseminated directly and efficiently to the educators.

The area of remote sensing, environmental management, GIS, and disaster management and mitigation requires education and training efforts to "think outside the box." Every effort to reach beyond traditional classroom paradigms to tap rich data and computational tools is important to the future strength of such efforts. Additionally, education and training efforts have to reach for resources well outside the physical confines of a single environment. Aggregation and prioritization education and training around science gateways the provide access to simulation tools and lower barrier of entry is a key and essential strategy.

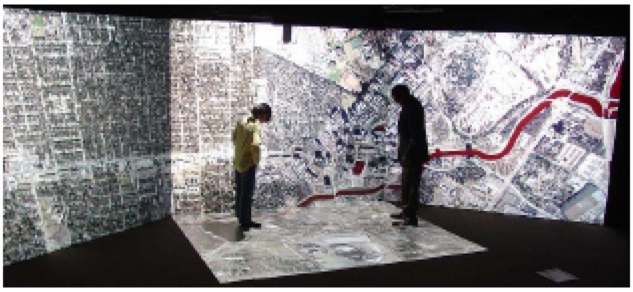
VISUALIZATION

In recent years, data from multiple sensors, real-time satellite data, NEXRAD Doppler Radar, hydrological instrumentation, and socio economic data has been increasing rapidly because of newer and faster methods of data collection and interpretation. Today, it is not uncommon to have spatial data sets with terabytes of data. However, the technology to process and display these kinds of data in real-time is still lagging. Traditional high-end systems to handle such large and complex data sets are very expensive.

A GIS system allows us to manage, analyze, and display geographic information which is represented as layers of information (ESRI-1, 2004). This information set includes – maps, globes, geographic data sets, processing and work flow models, data models, and metadata. The broad area of GIS can be viewed in different ways. For example (ESRI-2, 2004), as a spatial database of features, rasters, topologies, networks, etc. Or, as a set of intelligent maps with several views that shows features and feature relationships on the earth's surface to support queries, analysis, and editing of the information. Or, as a set of tools and functions to transform information that derives new geographic data sets from existing datasets. It is obvious that display of information is a very vital part of any GIS system. Large maps that are opened and spread flat give much more information than when they are seen folded. The more information we can see without loosing finer details, the better we can detect and identify objects and the better we are in making associations and connections between various features in a map. The human eye has the best image processor (Allan, 2004). We can make faster and better decisions than computers in many situations. This simple philosophy drives us to find better ways to visualize geographic data.

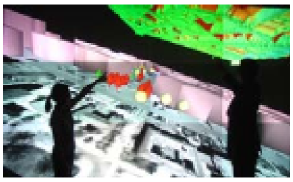
To process and visualize large data sets effectively requires combination of high-performance computing, smarter software algorithms, and large display systems. Traditionally desktop based graphics rendering is done using a single personal computer (PC) on a single monitor. With the latest developments in graphics hardware and computing power, a single PC is used to drive multiple displays at the same time. But the complexity of scientific visualization applications increased tremendously so that a single PC could not render complex scenes in real-time even to single display, not to mention multiple displays. This single PC, single and multi display configurations have several limitations, such as insufficient computing power, trade-off between rendering quality vs. display area, especially to render complex scenes in real-time. To overcome these limitations, we use commodity cluster to render complex scenes to a single monitor, and/or to different tiled display configurations in real-time. Our approach provides more flexibility and ability to customize the computing power, rendering quality and display area as required.

While the above systems provide high-performance computing, rendering and display area, most of the commercially available software cannot be used "as is" on these specialized hardware to its maximum potential. Most of the current software applications are limited and focused on solving specific problems. Our objective is to be able to run our visualization software on these high-performance systems seamlessly so that the user is able to visualize with the same ease as if it were running on a desktop workstation. The applications are in 3D and stereoscopic, and because they are rendered by a very powerful graphics cluster these displays are navigable in real-time. This software application not only takes advantage of distributed computing but also the distributed rendering for large, very high resolution display systems.



Visual result of flood simulation

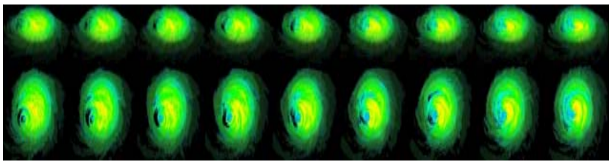
The image rendered on left is a visual result of a flood simulation. This flood simulation was developed using 1 foot color ortho images draped over a digital elevation model provided by 5-foot resolution LIDAR data. The data were provided by the Texas Advanced Computing Center (TACC) as part of a Flood Modeling demonstration to illustrate the promise of TerraGrid type activities hold for facilitating real-time simulation of emergency



Aerial visualization of a town; Layered view of multiple data response personnel and evacuation planning flood events over large regions. This was a joint effort involving TACC, Oakridge National Labs and Purdue University.

Figure above is visualization of the greater Lafayette region in Indiana, USA. The model shown consists of two sources of data: a 1m resolution color digital aerial photograph and a 30m resolution digital elevation model (DEM) of the area. ArcGIS software was used to drape the color image onto the DEM. Then the 3D draped model was then converted to VRML format and then to 3ds Max™ file format. This 3DS file was eventually used to display the model using inbuilt cluster aware GIS rendering application on the tiled wall display.

The above figure also shows the display of multiple layers of information of Purdue University Campus. They include – a one foot resolution aerial imagery of the Purdue University Campus, a layer of information with buildings, roads, sidewalks, and different species of trees, and a 1m resolution LIDAR data. The technique used to visualize this project was the same as discussed in previous application. Each layer was visualized in the ArcGIS 3D environment then passed to the cluster-aware GIS rendering application on the tiled wall display.



Timed Doppler Radar data Hurricane Katrina

Figure above shows time varying Doppler radar data of hurricane Katrina rendered in both top and front view.

The above described applications demonstrate the usefulness of visualizing complex data sets on large display systems. This gives us a very high resolution display that shows much more detailed information. While the software applications for visualizing individual aspects of these data sets are being developed, integrating all these data sets into a single system that coherently interrelates and provides an effective visualization interface for the users is still one of the biggest challenges.

Reference web pages
http://clearinghouse1.fgdc.gov/
http://edc.usgs.gov/products/satellite.html
http://daac.gsfc.nasa.gov/MODIS/software.shtml
http://registry.fgdc.gov/serverstatus/
http://www.fgdc.gov/nsdi/nsdi.html

REFERENCES

Allan, J. W. 2004. High resolution geographic imagery and its impact on GIS. *GIS Development*. 15 July 2004 http://www.gisdevelopment.net/technology/rs/techrs0015pf.htm.

ESRI-1 2004, GIS Concepts Overview - What is GIS? 12 July 2004. ESRI. 15 July 2004. http://esri.com/software/arcgis/concepts/overview.html.
ESRI-2 2004, Three Views of a GIS. 12 July 2004. ESRI. 15 July 2004. http://www.esri.com/software/arcgis/concepts/three-views.html.

Mackenzie, J.G., R.M. Allen, W.B. Earl, & L.A. Gilmour, 2001. Amoco computer simulation in chemical engineering education. *Journal of Engineering Education, 90*(3), 331 – 345.

Madhavan, K.P.C., Dutta-Bergman, M.J., & Arns, L.L., 2006. Pedagogical e-learning frameworks using advanced 3D visualization for bio-terror crises communication training. In S. Amass, A. Chaturvedi, and S. Peeta. (eds.) *Advances in Homeland Security. Vol. 2. Guiding Future Homeland Security Policy - Directions for Scientific Inquiry*. West Lafayette, IN: Purdue University Press. pp. 33 – 53.

Oroda, A. S. 2001.Application of remote sensing to early warning for food security and environmental monitoring in the horn of Africa. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34, 66- 72

Parent, Florence, MD, MPH; Yves Coppieters, MD, MPH; Marc Parent, MD. Information Technologies, Health, and Globalization: Anyone Excluded? *Journal of Medical Internet Research*. 3, e11, 2001.

Pfitzer S, Verwoerd D, Gerdes GH, Labuschagne AE, Erasmus A, Manvell R, Grund C. 2000. Newcastle disease and avian influenza A virus in wild waterfowl in South Africa. *Avian Diseases.* ;44, 655-60.

Reigeluth, C.M., and E. Schwartz. 1989. An instructional theory for the design of computer-based simulations. *Journal of Computer-based Instruction, 16*, 1 – 10.

Samara, J. S. 2004. Review and Analysis of drought monitoring, declaration and management in India. *International Water Management Institute.*Working paper 84.

Schreppel, C. K., 2003. Setting the alarm for an early warning. *AWWA* . 29(6): pp.7

Skidmore A. K., W. Bijker, K. Schmidt and L. Kumar.1997. Use of remote sensing and GIS for sustainable land management. *ITC Journal*,:302 – 315

Skyttner, L., 2002. Monitoring and early warning systems – a design for human survival. *Kybernetes*, 31,220–245.

UNCCD, 2003. Early warning systems. The United Nations Convention to Combat Desertification (UNCCD). 2000 Königswinter, Bonn, Germany (pp 8-12) 2001, Fuji Yoshida, Yamanashi, Japan. (18-25) Verdin, J.,C. Funk, G. Senay, and R.

Vieth, T.L., J.E. Jobza, and C.P. Koeling, 1998. World Wide Web-based simulation. *International Journal of Engineering Education*, *14*, 316 – 321.

Wessels K.J., S.D. Prince, and J. Small. 2003. Monitoring land degradation in southern Africa based on net primary productivity. IEEE 5:3305 – 3307