We need **Scientific** Data Management!

This is a story inspired by a recent meeting of the Community/Council for Data Integration (CDI), held at the Denver Federal Center, August 10-12, 2010.

I recall that Kevin Gallagher asked for people in the community to collect "stories", presumably about CDI and for the benefit of others in the community. After the meeting, we were sent links from Heather Henkel and Katherine Lins, and they discussed issues of data mining and scientific integrity. This story is more personal. I will try to explain why I think Scientific Data Management is worthy of our best efforts. Scientific data management is very different from operating an archive, storing and allowing data access through a controlled environment, or even periodically reading and rewriting the files to maintain their readability. This is not enough.

Who I am - my training

I began my career in California back in 1970, as an undergraduate at the University of Redlands. It was the spring semester of my junior year, when I had a unique opportunity to work as a full-time intern with a 24 year old fellow who was just out of graduate school and starting a new company. His name was and is Jack (yes, that Jack). He introduced me to spatial analysis, the software tools we now call GIS. After that academic semester, I joined the small staff at ESRI, where I worked as a project manager and later as the "Coordinator for Technical Services" with projects ranging from building a database for the USFS Tahoe National Forest, the State of Maryland (MAGI), Texas Water Development Board (Automap II), Takanaka Construction – new town development in Japan, to numerous small contracts with architects and engineers. We worked long hours and generally had a great time. I worked with Jack at ESRI until 1974, when I decided to move my family back to the upper Midwest. There I began work at a new facility called the USGS EROS Data Center. At EROS, I learned about satellite remote sensing and they sent me to Purdue University for specialized training in digital image processing of remotely sensed data. I worked to integrate my GIS skills with remote sensing applications in the Data Analysis Lab. Along the way, I had the privilege to work with USGS geologists, biologists, and hydrologists (e.g. Bill Fischer, Virginia Carter, Richie Williams, and Chuck Robinove). For more information about the time from 1970 to 1985, read Chapter 11 of "The History of GIS", edited by Tim Foresman, Prentice Hall, 1998. Steve Guptill and I contributed this chapter to the book that describes the activities in DOI. Without a doubt, I was lucky to have had the opportunity to team up with some world class scientists as they worked on the problems of the day. As a result of these experiences, I am thoroughly hooked on the value of geography and GIS for organizing data and studying big problems.

Introduction to Scientific Data Management

In the EROS Data Analysis Lab, we kept complete files for each of the scientists and documented their projects. All this was done in parallel with the scientist and his/her documentation. The files were useful if they ever requested information about their work in our lab. Over time, some of these people have retired and some have died. Over time, we have gone back through the files to consolidate and purge the nonessential items.

The Great Flood of 1993 - SAST

In 1993, there was a major flood on the Upper Mississippi River and the Lower Missouri River that resulted in the selection of an interdisciplinary group of 18 scientists called the Scientific Assessment and Strategy Team (SAST). EROS supported this team with scientific and technical support, and that was my job. By winter 1993/1994, the team had assembled a large database of digital imagery and data gathered from many agencies, some of it from maps that we digitized, and much of it that required documentation (metadata) that kept 2 full time specialists busy working along with the team. Several million dollars of high resolution lidar topography data was collected to document the intricate details of scour and deposition in the river channel. SAST wasn't just about publications and notes – it was original data and imagery that documented this flood – truly a "snapshot in time". We developed a "levee database" from maps and data provided by several US Army Corps district offices. This was later turned over to the Corps for them to extend and maintain. When it came time to archive the SAST data, the EROS archive didn't have a place for it, other than set up a rack off to the side of the satellite imagery archive. This is where everything sits, in boxes and map tubes. In 1995, we set up a server, and made much of the digital SAST database available on the fledgling Internet. A real coup! Three servers later, with various upgrades to metadata standards, relational database engines, and numerous releases of GIS software, database structures, and file formats, I continue to get requests for SAST data, even though the money ran out in 1997. Also, I get regular encouragement from the SAST project leader who is now at Penn State. A few years ago, the EROS archive staff proposed and was funded to perform a modest "data rescue" effort to redocument the SAST archive, using a summer hire and modern software tools. Everything helps.

ESRI CRADA - Seamless Server

In 1998, USGS and ESRI began work on a Cooperative Research and Development Activity (CRADA) to create a method to deliver seamless elevation data and derivatives, as well as imagery. ESRI retooled their Spatial Database Engine to handle raster data, and USGS finished up the first time coverage of cleaned up elevation data that was called NED – the National Elevation Dataset. In 2000, we together launched the Geography Network at the ESRI International Conference. If it is true that we all get 15 minutes of glory, then 4 of my minutes will come from the plenary session of that conference, where we demonstrated our server to 12,000 people. Shortly thereafter, the EROS Director reassigned me to the Data Services Branch to take our cooperative research and move it to production – and then he proceeded to retire! I struggled to get something done with no management support. In those days, Data Services sold lots of data products and brought in millions of dollars that they could turn around and spend. Even though OMB Circular 130 clearly states that agencies may not recover more than the cost of reproduction, this was a bit too tempting for a fox in the chicken coop. My ideas didn't fit their business model. I became "Distribution Dave", the guy who wants to "give away our data". Fifteen years later, HQ instructed EROS to "web enable" all the Landsat data, and now we can all celebrate the new users and the new applications of the data in the archive.

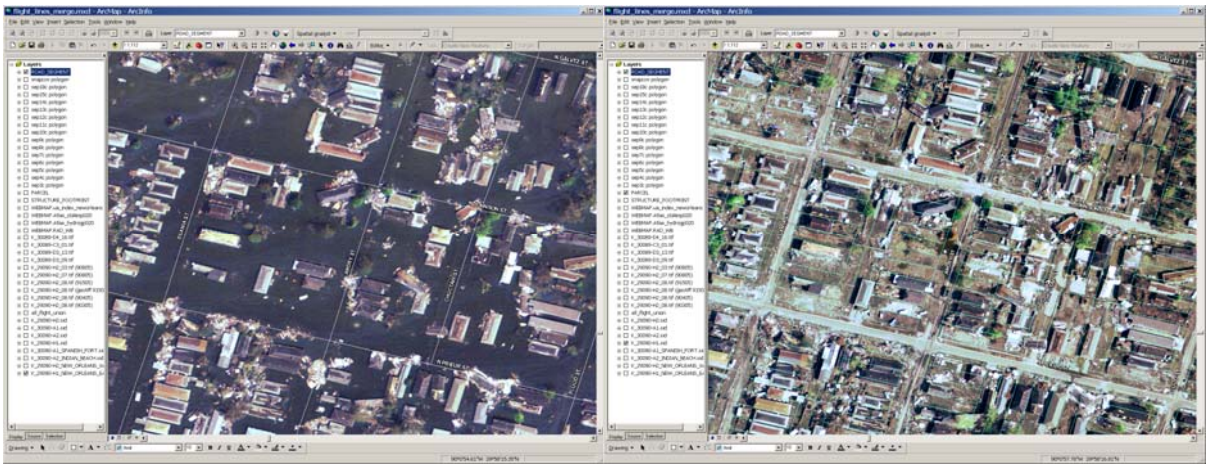
Katrina – GIS for the Gulf

Soon after Katrina hit shore in 2005, I got a phone call from a USGS colleague. USGS and ESRI were going to work together to collect GIS data, down to building footprints and parcel boundaries, and to do it for hundreds of counties and parishes in four affected states. I went to California for about 3 weeks, and so did 4 hard working colleagues from Rolla, MO, Denver, CO, Menlo Park, CA, and Sioux Falls, SD. We worked side by side with ESRI staff, using new "extract, transform, and load" (ETL) tools that establish data models for ingesting local data to the "GIS for the Gulf" database. We sent datasets to ESRI staff who were embedded in the Joint Field Offices (JFOs) in each of the affected states. We also kept our USGS managers informed, but our efforts were apparently not considered mainstream USGS science contributions, and were, for reasons that I don't fully understand, largely swept under the rug.

A story within the Katrina story

Within a few days after we started collecting local GIS data, we began getting imagery from the USGS emergency response folks at EROS. Every few days, more imagery would come online or came to us on hard drive. Over time, I noticed that some of the newer deliveries replaced existing imagery. As I looked at the flight line shape files that were also delivered with the imagery, I realized that the contractor had flown the area over a period of almost 3 weeks, and some areas had 4 or even 5 repeats. In the end, there was a "final delivery" and it was generally the latest look, but the one least likely to be during the peak of the flood. Apparently they had trouble with the geometry, so that some areas were reflown to meet spec. When I got back to EROS, I documented the chronology from the individual flight line shapefiles, and showed examples of identical file names with scene content as much as 2 or more weeks later. EROS data management had decided to store the final delivery, and disregard and effectively discard the earlier imagery. I argued vehemently, but to no avail. I view this decision as a blunder bordering on scientific misconduct. I kept the earlier imagery under my desk, and on several occasions tried to revisit the decision. After we all got an email telling us to not discard anything that pertained to Katrina, I was even more certain that we should go back and do the right thing to document and reinstantiate the earlier imagery.

Below are examples of Katrina imagery acquired on September 3 (on left, discarded). On the right is imagery for the same area acquired on September 15 (final delivery):



In the final delivery, metadata indicated that the imagery had been flown some time in the month of September, and yet the flight lines recorded all the dates and times. We should have connected the two (and we still should). The capper for me is that I was disinvited to a meeting on Katrina "lessons learned", presumably to keep me from telling the story of GIS for the Gulf and the discarded imagery. There are a couple people that I lost considerable respect for during this experience, and they know who they are.

How should Scientific Data Management be done?
In my opinion, and in 40 years of experience, I believe that the archive and management of scientific data and results should be done by people who understand the applications of the data, and not by those who simply treat is as a librarian, carefully logging it, protecting if from the elements, and maybe even freshening it by copying it to new media every few years. This is just not enough. Scientists and users of the data will see a need to upgrade to new searchable metadata formats, to further improve and process the data with newer tools, while making certain that no data is lost or rendered unusable because of obsolescence or irreversible data reductions. At the CDI meeting last week, John Caron from UCAR told us that they pride themselves on "eating their own dog food", with the implication that they know what is important because they use their own data. This is in contrast to what we sometimes hear – that we tend to "throw it over the wall". I prefer to say that GIS data looks entirely different to a person who attempts to use it – and this should guide database design, archive, and management. A complication is that the funding for most scientific projects is generally of the 1 or 2 year variety, and that the scientific data management job extends far beyond that – some people say that the job is perpetual – now that's a long time!

What about CDI and this new community of interest?
Scientists should not just hand everything to a librarian or an archivist and then walk away from it. If some scientists do that (walk away), that's not the way we (USGS) should leave it. Some aspects of these databases may need to be periodically updated, translated to new structures, data formats, and even reprocessed to assure that they will be useful for future generations. We need to maintain and enhance linkages between databases, new and old. Imagery archives have their own challenges associated with media shelf life and sheer volume, but scientific data management should be done by scientists –by those who truly understand the data. They should have a working knowledge of interoperability standards and access to the best tools of the trade.

In the time since I began cranking out paper tape digitizer records at ESRI or began pushing pixels at EROS, much has changed in terms of our ability to analyze the datasets and to get accurate results. Some of these old projects should be revisited from a historic perspective, and databases should be carefully retooled. Geologists might say that we should "remine the tailings" for their historic value.

Some examples are: 1) LUDA/GIRAS - USGS spent considerable time and effort documenting land use and land cover during the 1970's. EPA took on the custodianship and delivered the data for many years. Recently, WRD (Curtis Price) has cleaned up the data and converted it to modern data structures. Is there more we should do to facilitate

the measurement of change? We should pull together the people who know this database before they all leave and we lose their "corporate memory". 2) We should stage the EROS archives of historic imagery not just to be accessible or web enabled, but truly easy to use. ESRI has offered to help. They say they want to make their software more "image aware", by allowing easy access to archived imagery. We should support that effort and work as full partners to make that happen.

Many good scientists manage their own data, even after funding is gone. They do this because they understand its true value. Also, they do it because it's the right thing to do. I met some new ones (good scientists) last week at the meeting in Denver. We need to make certain that their good work is not lost or forgotten when they are gone. We need scientific data management.

Dave Greenlee
8/19/2010