Building a System to Disseminate Digital Map and Geospatial Data Online

Tsering Wangyal Shawa

Abstract

The expectation of library patrons to get all of the information they need, including geographic information, accessible on their desktops has created challenges to map and Geographic Information System (GIS) libraries. This new expectation has forced libraries to think about how to design a system that will allow diverse geographical information to be available over the Internet. Some libraries have built a site to distribute localized data, others have developed a system to make only maps accessible online. Princeton University Library's Digital Map and Geospatial Information Center started a pilot map scanning project in early 2004 to build a system, to develop specifications for scanning maps and compressing TIFF images to JPEG2000 file format, and to establish workflows. The system was built using many off-the-shelf commercial software packages. This article discusses challenges of building a system and explains how Princeton developed a scanning process and standards, workflows, and what lessons were learned in building such a system.

Introduction

Libraries purchase and receive geospatial data and paper maps free of charge through the Federal Depository Library Program (FDLP). One of the requirements of the FDLP is to make all the materials distributed through it freely accessible to the public. Because of this requirement and demands from library users to make all the materials accessible on their desktops, many libraries scan their paper maps and make them accessible online. However, one major problem libraries face is how to design a system that will allow the user to search, view, and download diverse geospatial data

LIBRARY TRENDS, Vol. 55, No. 2, Fall 2006 ("Geographic Information Systems and Libraries," edited by Jaime Stoltenberg and Abraham Parrish), pp. 254–263.

and digital maps. This article examines the challenges of creating such a system and explains how Princeton University Library's Digital Map and Geospatial Information Center has designed a system that will allow the library to integrate various forms of geographic information and make them accessible online from one interface.

Challenges

There are numerous challenges in making geospatial data and digital maps accessible over the Internet. Many libraries have used ESRI's ArcIMS and ArcSDE, and relational databases such as Micosoft's SQL Server, Oracle, etc., but they were not very successful in making diverse collections of digital maps and geospatial data accessible online from one interface. This was due to the following reasons:

- Disseminating digital maps and geospatial data via ArcIMS technology is not practical for libraries when they have a great quantity of material covering different parts of the world at different scales and in different formats.
- There is no simple way to view and download vector geospatial data stored in ArcSDE without creating ArcIMS image or feature services. Using ArcIMS to build image and feature services to view and download vector data is not only time consuming but also uses a lot of processing power on a server.
- Many libraries are scanning large historical maps and aerial photographs. Some of them are georeferenced but many are not. Disseminating these types of materials with vector geospatial data is a real challenge.
- The file sizes of scanned maps and geospatial data could vary from a few megabytes to a gigabyte. Making a large file accessible over the Internet is a challenge.
- Designing a system that has easy workflows and ease of maintenance is difficult.

Because of these reasons, I spent a few years testing different server side technologies to build a system that will not only allow our library to organize and manage digital maps and geospatial data with easy workflows but will also allow users to search, browse, view, and download different formats of geographic information. Some of these formats include scanned historical/present maps, aerial photographs, satellite images, and vector geospatial data. The advantage to building such a system is that all kinds of geographic information can be integrated, managed, searched, and accessed from one interface. Geographic information can range from maps and geospatial data to photographs of places, etc. Many libraries have designed systems to disseminate maps and geographic data online, but the focus is either regional or item specific. In order to build an integrated system to disseminate diverse geographic information, I started a pilot map

scanning project in early 2004. The goal of the project was to design systems and specifications for scanning maps and to establish workflows.

System Design

Before designing a system I had to research what kinds of software packages were available. The Environmental Systems Research Institute (ESRI) server software packages were some of the most sophisticated software packages on the market and some of the most easily available to academic institutions because of ESRI educational licenses. The ESRI server software packages could handle most of the things that I wanted to accomplish. For instance, storing data in ArcSDE provides the flexibility to make data accessible to ArcMap users over the Internet and to store data in a relational database management system (RDBMS). However, there are some limitations to the software. The ESRI server software packages assume that all the data will be made accessible online via ArcIMS and will be georeferenced. That leaves out all the scanned maps or aerial photographs that have no georeferenced information. Another limitation with the ESRI software is that if data are stored in ArcSDE, the only way for a non–ESRI software user to access these data over the Internet is to build some sort of ArcIMS service and make it viewable and downloadable in shapefile format. This server design forced me to look for different software packages that offer the ability to disseminate non-georeferenced scanned maps and aerial photographs online and provide users with the option to view and download vector data straight from ArcSDE.

After understanding the pros and cons of using ESRI server packages, I built a system using ESRI server software packages such as ArcIMS MetadataServer, ArcSDE, Micosoft's SQL Server database, and ArcCatalog. I also used off-the-shelf commercial software packages such as Safe Company's SpatialDirect/FME and Mapping Science's GeoJP2 Encoder and Decoder and Image Server. I used ArcCatalog to create metadata; ArcIMS MetadataServer, ArcSDE, and SQL Server to publish and store all the metadata and geospatial vector data; and SpatialDirect and FME to access data from ArcSDE and convert ArcSDE data into more than thirty different file formats. I used GeoJP2 Encoder to convert and compress TIFF files to JPEG2000 (JP2) and Image Server to serve JP2 images over the Internet without plug-ins. I was able to create five databases (Metadata, Gazetteer, GISdata, SpatialDirect, and PUMapData) in the SQL server to store various components of our data. The Metadata database stores all the metadata records, the Gazetteer stores gazetteer information to help search a place name more easily, GISdata stores all the vector data, SpatialDirect stores all the vector records to interact with FME software, and PUMapData stores basic information of scanned maps and creates unique image file names. In addition to these databases, I also created two folders in our server to store JP2 images. One is for holding public domain materials, and the other is

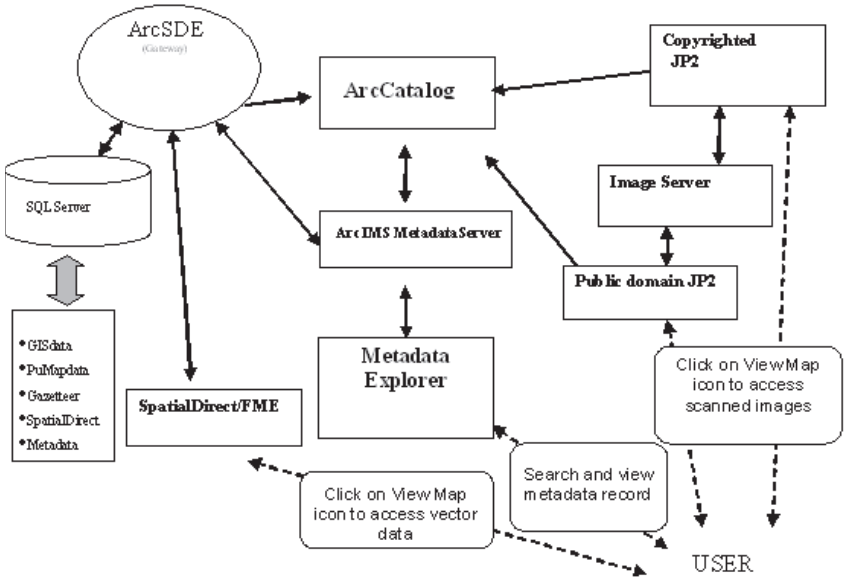

Figure 1. System Design

for storing copyrighted maps. Both of the folders are linked to JP2 Image Server. See Figure 1 for a diagram of this system.

Scanning Process

Before the scanning work was started, I researched how other institutions were scanning maps and why specific resolutions were used. The Library of Congress scans cartographic materials at 300 dots per inch (dpi) with tonal resolution of 24-bit color and saves files in TIFF non-compressed file format. The British Ordnance Survey (OS) scans maps between 254 dpi and 400 dpi in a non-compressed TIFF file with 256 colors. The United States Geological Survey (USGS) has done a lot of map scanning work. The main goal of the OS and USGS scanning work is to convert paper map information into digital geospatial data. The USGS has scanned differently scaled USGS maps, extracted map information, and created geospatial data such as digital elevation models (DEMs), digital line graphs (DLGs), and U.S. Census's TIGER street data. In researching what resolutions were used, I found that the USGS Digital Raster Graphics (DRGs) made before October 2001 were scanned at the resolution of 250 dpi. However, most DRGs made after October 2001 have scan resolutions of 500 dpi. The colors of the scanned maps were reduced to a standard color map of 13 colors. The goal of most map scanning projects is to preserve map information and extract data from the map for geospatial analysis.

An in-house test proved that scanning a paper map (USGS 1:24,000 topographic map) at 400 dpi with 256 colors versus 500 dpi with 24-bit color shows very little difference. In fact, most of the large-format sheet-fed scanners that are currently on the market have around 400 dpi as actual/optical scanning resolutions. Scanning a map higher than the scanner's optical resolution is basically interpolating actual optical resolution, which means the number of pixels and file size increase but better map information is not necessarily captured. After reading about and testing different scanning options, I came to the conclusion that a minor visual quality improvement hardly justifies the larger file sizes (500 dpi with 24-bit color: file size 441MB; 400 dpi with 24-bit color: file size 278MB; 400 dpi with 256 color: file size 96.2MB). Nor does it justify the extra time it takes to scan and save the image. Therefore, I decided to scan paper maps at 400 dpi optical resolution with 256 colors, since scanning a map to preserve map information for later Geographic Information Systems (GIS) use and scanning a map as artwork are two different things. The objective of this scanning project was to preserve map information, so it was not important to capture all the subtle color differences or color "noise" generated by the condition of the paper and the printer. Maps published by the USGS usually use less than 13 colors, and storing a scanned map as 256 colors is more than enough to preserve map information.

After making the decision on what resolution to scan the maps, I also needed to research what was the best compression ratio to encode the TIFF file into JP2 file format. By performing different compression tests I found that 10:1 was the best compression ratio in terms of visual result and file size. The maps were scanned at 400 dpi with 256 colors and were saved in a non-compressed TIFF file format for archival purposes. The TIFF images were then compressed using GeoJP2 software into JP2 files with 10:1 compression ratio for online access.

Once scanning resolution and compression ratio standards were established, the maps were scanned without making much effort in color balancing, image cleaning, or other changes in image processing software. One exception to this was that the images were cropped to delete white space that was not part of the map. Any pencil marks on a map were erased before it was scanned. In the initial stage, our library scanned maps covering different parts of the world to organize them in different geographical regions and to test how browsing options worked on the Metadata Explorer's page.

Workflow

The maps scanned as part of this project were cataloged in the GEOMAP database (our local map cataloging database). Before a map was scanned, the catalog record was located in the GEOMAP database and used to enter brief information in the PUMapData database. A simple Microsoft Access

interface was used to connect to the PUMapData database, which is located in the SQL server. Once a connection was made, a staff member entered brief information about the scanned map, such as the title, publication date, and description of how the map was scanned and encoded, etc., in the PUMapData database. After entering the basic information, the database allowed us to generate a sample text file consisting of the information entered in the database along with a unique ID and the time and date the map was scanned. This was used as a brief metadata record and was encapsulated with the scanned map when it was encoded into the JP2 file. The unique ID was also used as a file name for the scanned map. The scanned map was saved as a non-compressed TIFF file. Afterwards, all the scanned maps were compressed (encoded) with text generated from the PUMapData database, using Mapping Science's GeoJP2 Encoder software. Once the maps were compressed, they were moved to JP2 folders in our server. The public domain maps were moved to a normal JP2 folder. If the scanned map was copyrighted, it was moved to another folder called "Copyrighted." The maps from this folder are accessible only at one computer in the Map Library. The non-compressed TIFF files were moved to a specially designated hard drive space for archiving.

Once maps were in the JP2 Image Server folders, metadata records were created with ArcCatalog software. All the scanned maps were individually cataloged using the *International Organization for Standardization (*ISO) 19115 metadata standards. At this stage, the GEOMAP database was accessed in order to pull the compressed map catalog record using a GN number (all the scanned maps that were cataloged in GEOMAP database have this unique number). Most of the GEOMAP catalog record is used for creating metadata for scanned maps in the ArcCatalog. Once a metadata record is created, it is published to the ArcIMS MetadataServer. As soon as metadata is published, a scanned map is immediately accessible to our users. Before publishing metadata, we created different folders in the MetadataServer that are based on some geographical hierarchy such as continents, regions, etc. (for example, North America ® United States ® New Jersey ® Mercer County ® Princeton). These folders are used for publishing our metadata records and will help our users to browse and select a map based on some well-known geography hierarchy.

After publishing the metadata, the scanned map ID and name were entered in the Excel spreadsheet with a note stating the metadata record was created. If somehow a metadata record could not be created or there was a problem with a compressed image, that information was entered in an Excel spreadsheet for a substitute record.

Vector data workflow processes are slightly different. First the data were uploaded in the ArcSDE using ArcCatalog, and SpatialDirect's Spatial Assistant connected ArcSDE tables (this connection allows SpatialDirect to read the data directly from ArcSDE without creating ArcIMS services). After

making the connection between ArcSDE and SpatialDirect, we opened SpatialDirect's Administration Interface Web page, created a map image, generated a unique URL, and entered the necessary information such as file name and size in the database called SpatialDirect. This database is located on the SQL server. We then opened ArcCatalog and made a connection to the ArcSDE database. We selected data and created a metadata record for that data, and while creating the record we inserted a unique URL that was generated in SpatialDirect in the Online Linkage space. Next we saved the metadata record and published it in the ArcIMS metadata server. The published metadata and data were then ready to search, view, browse, and download from Metadata Explorer immediately. Figure 2 shows a snapshot of a Metadata Explorer page.

How the System Works

This system helped the library develop an easy workflow and also helped patrons search and browse geographical information including geospatial data, maps, and aerial photographs from one interface without searching different databases. The system has also allowed our library to scan copyrighted maps in addition to those in the public domain. Copyrighted maps are scanned for two purposes: for archival reasons and to give a general picture of how a map looks. This is possible because the scanned materials that have metadata records also have thumbnail images of the map. This thumbnail image of the map will give our user some idea of whether the map in our library will be useful for his/her research. If we did not provide this option, users would need to come to the Map Library to look at the maps.

This system design has given our patrons the option of accessing our materials on their desktops, either by searching or by browsing. Once the material is found, a user can click on the *View Map* Icon to view a map as a digital image or vector data. If it is a public domain map, the user can view and download the map in either JPEG or TIFF. If the map is georeferenced, the user can not only view the map but can also download it in JPEG and TIFF with a world file. This allows patrons to use a downloaded map in GIS software. If it is a copyrighted map, then another window will pop up with a message stating it is copyrighted material and that the map is not accessible over the Internet but can be viewed at the Map Library.

If the user is accessing vector data, the system will force the user to type his/her user name and password. User names and passwords are necessary to protect misuse of SpatialDirect/FME software. These software packages are free for academic institutions for educational use, but the Safe Company does not allow use of the software by the general public. Once the proper information is provided, a general coverage of the map will be shown, which allows the patron to download the file in more than thirty different

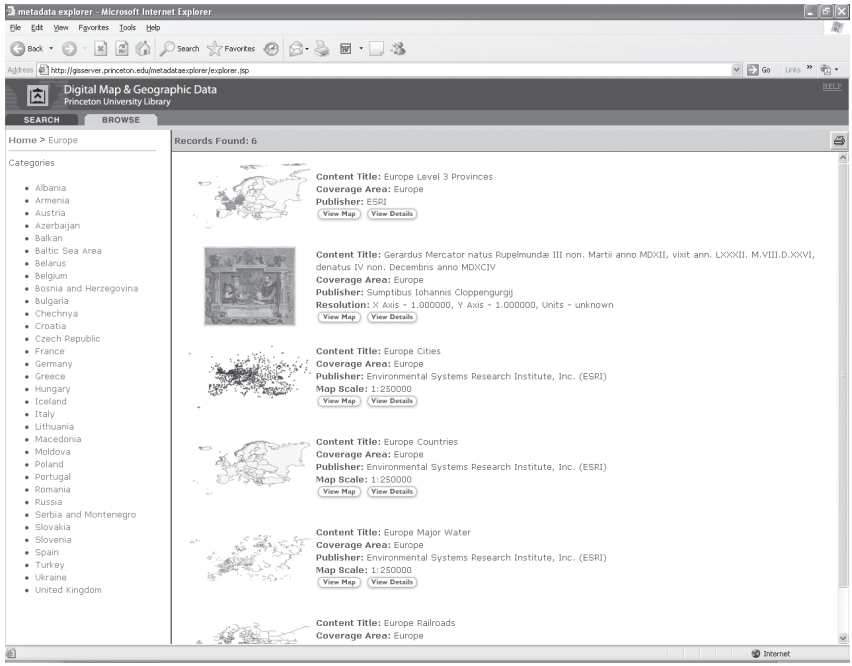

Figure 2. Sample Metadata Explorer Page

file formats. This has allowed our users, who may not use ESRI software, to access and download data in their preferred software file formats.

Lessons Learned

Building a digital data infrastructure has helped me to understand what resources are needed and how to build such a system. I found that it is crucial to get support from the systems department, a database specialist, and a programmer to design a system; without their help it would be very difficult to build and maintain a system. To continue with scanning, creation of metadata, and uploading of vector data in ArcSDE, having a dedicated support staff is essential. Based on these experiences, we have found that hiring student workers may not be the best option. The high turnover among student workers every semester demands too much time and resources for training. This high turnover can also lead to inconsistent quality of work.

Throughout this project, we found it was important to make the library administration understand what size of disk space we needed for our work. After I was initially given a server with roughly 300 GB space, I informed

our administrator that this was not enough. I suggested a minimum of a few terabytes of server space to continue with our map scanning project and making geospatial data accessible online. Unlike other digital projects, scanned maps and geospatial data take up a lot of disk space, and therefore it is important for the library administration to understand the need for the larger amounts of disk space to continue with the work. In addition to disk space, I also learned from my experience the importance of building a redundancy system on our server so that if anything unexpected happens, our services will remain accessible to our users. Because of this, we decided to move our system to a new server that is based on a cluster server. This server has two nodes, both of which will be running the same application but data will be stored in another drive. This server design will help us to build a redundancy system. The final lesson that I learned was the need to create an alias name for the server. This way, when we move the project to another server, we can keep the same alias server name and will not have to change the Web page address/name.

Conclusion

The pilot map scanning project was very helpful to our library. It helped us build a system that will allow our library the flexibility of disseminating diverse geographic information over the Internet. Before the system was built we did not have the tools to make maps, aerial photographs, and geospatial data accessible online from one interface. The project allowed us to use a new file format called JP2 and to develop our map scanning and file compression standards, which we continue to use. It helped us to estimate the size of disk space we need to continue making our diverse geographic information available to our library users online. It also helped me make our administrator aware of what supports and resources were needed to integrate diverse collections of geographic information and make them accessible online. One of the goals in designing this system was to encourage other libraries to build similar systems for their own use. In addition, the project led me to ask the president of ESRI to develop a similar system for the map and GIS library community. If ESRI does design such a system, my hope is that it will minimize the complexity I found in integrating different software packages. Whether libraries manage to build their own systems or are able to use a new package from ESRI (if they do design such as system), I hope that more libraries will be encouraged to make their diverse geographic data accessible online from one interface.

Appendix: Suggested Reading

British Ordnance Survey. (n.d.). *1:25 000 Scale Colour Raster: technical information*. Retrieved November 12, 2003, from http://www.ordnancesurvey.co.uk/oswebsite/products/25kraster/techinfo.html.

British Ordnance Survey. (n.d.). *1:10 000 Scale Raster, technical information*. Retrieved November 12, 2003, from http://www.ordnancesurvey.co.uk/oswebsite/products/10kraster/techinfo.html#gr.

GPO. (2005). *About the Federal Depository Library Program (FDLP)*. Retrieved February 8, 2006, from http://www.gpoaccess.gov/fdlp.html.

Library of Congress. (n.d.). *Scanning cartographic materials*. Retrieved November 12, 2003, from http://memory.loc.gov/ammem/gmdhtml/gmddigit.html.

Shawa, T. W. (2003). Review of JPEG2000 and GeoJP2 Compression Software. *Baseline: A Newsletter of the Map and Geography Round Table, 24*(3), 8–10.

Shawa, T. W. (2003). What is the best resolution to scan a map? *Baseline: A Newsletter of the Map and Geography Round Table, 24*(6), 6.

Shawa, T. W. (2005) From the chair. *Baseline: A Newsletter of the Map and Geography Round Table, 26*(5), 4–5.

USGS. (2001). *National mapping program technical instructions, standards for digital raster graphics*. Retrieved November 12, 2003, from http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/drg_temp/Pdrg0401.pdf.

USGS. (2001). *National mapping program technical instructions, part 1, general standards for digital raster graphics*. Retrieved November 12, 2003, from http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/drg_temp/1drg0401.pdf.

USGS. (2001). *National mapping program technical instructions, part 2, specifications, standards for digital raster graphics*. Retrieved November 12, 2003, from http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/drg_temp/2drg0401.pdf.

Tsering Wangyal Shawa is a Geographic Information Systems Librarian at Princeton University. He has widespread experience in geospatial data selection, software, and hardware and holds degrees in the areas of library science, education, geography, and cartography. He is the current chair of the American Library Association Map and Geography Round Table (2005–2006) and is the chair of the Geographic Technologies Committee. He was selected by the National Research Council and the Federal Geographic Data Committee's Homeland Security Working Group to study and publish reports on "Licensing Geographic Data and Services" and "Guidelines for Providing Appropriate Access to Geospatial Data in Response to Security Concerns." He was a consultant for the Tibetan and Himalayan Digital Library (THDL) based at the University of Virginia and was a Cartographic Users Advisory Council (CUAC) member from 2002 to 2005. He was born in Tibet and has lived and taught geography and cartography to high school and college students in India, Nepal, Kenya, and Sudan.