The Challenges and Benefits of Distributing Digital Data: Lessons Learned

By Kenneth R. Papp[1], Susan Seitz[1], Larry Freeman[1], and Carrie Browne[2]

[1]Alaska Division of Geological & Geophysical Surveys
3354 College Rd.
Fairbanks, AK 99709-3707
Telephone: (907) 451-5039
Fax: (907) 451-5050
e-mail: {ken.papp, susan.seitz, larry.freeman}@alaska.gov

[2]Formerly with the Division of Geological & Geophysical Surveys

INTRODUCTION

The Alaska Division of Geological & Geophysical Surveys (DGGS) has been producing geologic maps using a Geographic Information System (GIS) since 1983 (Davidson, 1998). The maps, reports, and informational publications produced by the DGGS are widely utilized by oil, mining, and resource-based companies, as well as consultants, universities, schools, government agencies, scientists, and private individuals. These users have become more technologically savvy over time, and as a result, user requests for digital data in addition to or in lieu of paper reports have grown exponentially. Since 1983, the DGGS has provided several web-based digital geologic data-distribution tools to accommodate the needs of its users (DGGS Staff, 2005), including a database-driven publications query page (http://www.dggs.dnr.state.ak.us/pubs/pubs.jsp), a geochemistry search engine (http://www.dggs.dnr.state.ak.us/webgeochem/index.jsp), and a geologic map indexer (http://maps.akgeology.info/). Currently, the DGGS provides users with digital versions of its maps and reports as Portable Document Format (PDF) files. The raw digital data that generate each map can be burned to a CD or DVD and purchased for a small fee through a general order process. At the time of this writing, a focused effort is underway to upgrade the DGGS web site to provide users with the digital data[1] used to create the state survey's geological and geophysical maps. This paper discusses the challenges and benefits of distributing digital data on-line.

PROJECT BACKGROUND

The primary goal of digital data distribution is to make the data available to the widest possible user audience in formats that are easily adaptable to typical end-user systems. The DGGS Digital Data Distribution (D3) Project is the result and distillation of a series of discussions convened in 2005 in response to numerous public requests for digital geologic map data and the need to fulfill digital data delivery requirements of some projects. The project provides end-users with a means by which to acquire all of the digital datasets used to create, "on the fly," DGGS geological and geophysical maps in the form of ESRI shape files, raster images (i.e. GeoTIFF), various other data, and metadata as compressed files for download via the World Wide Web. The scope of the D3 project includes the following: (1) Enhance the current publications pages to distribute compressed digital data packages, (2) develop a secure, internal application that will allow DGGS staff to create, on-line and off-line, packages for distribution to the public, (3) develop a means by which to insert metadata file elements into the database, (4) modify database structures as needed for application design, and (5) write accurate documentation for project process steps and changes made to the database.

The process work-flow for the distribution process is shown in Figure 1. The first step will mainly involve "cleaning up" the GIS data, ensuring that each dataset has valid metadata, and loading the metadata into the DGGS database (Freeman, 2001a, 2001b). Once this step is completed, all of the digital data files that comprise each publication (project file) will be archived and indexed into the DGGS database system, creating a record of the subsequent distributable dataset. Steps 2 – 7 will be accomplished by providing the authors with a secure, internal, web-based application that will allow users to index their digital geospatial data files and organize them into "packages" according to publication number, dataset, and then data format type. The GIS Manager will then review the dataset packages for data quality purposes (Step 7). Finally, the distribution package will be published to the

[1]Note: Underlined words are defined in Appendix A.

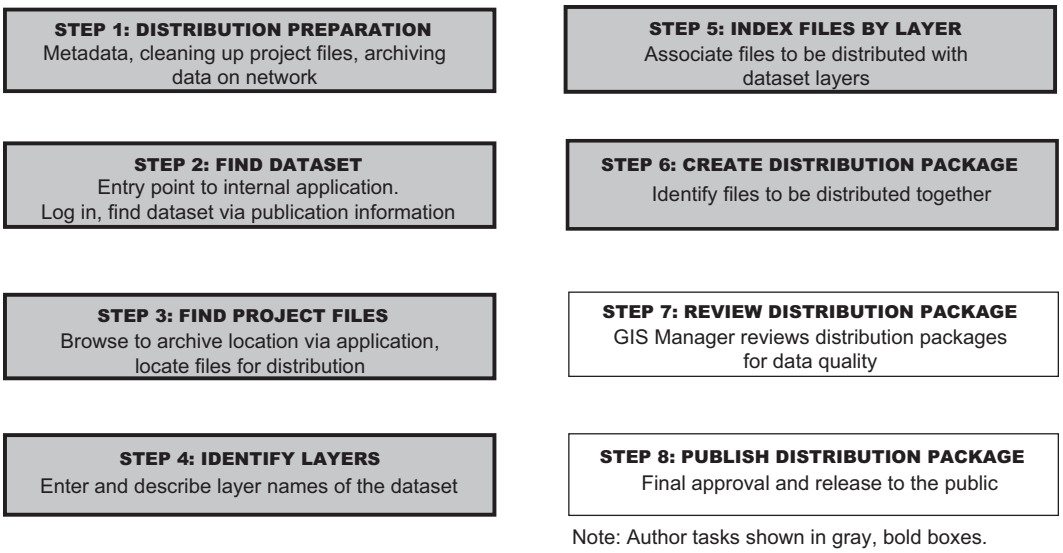

Figure 1. General workflow diagram for the DGGS data distribution process.

DGGS Web site and available to a wide array of end users upon final approval by the DGGS Director (Step 8).

PROBLEMS THAT NEED SOLVING

The old saying "Garbage In, Garbage Out" is certainly relevant in this case, and the process of finding and cleaning up the raw digital data can be daunting and consume a large portion of the D3 project's resources. Once the data files for publications have been located, recovered, updated, and stored in a central location, a decision needs to be made regarding the data format that will be distributed to the end user. A project of this type and magnitude also requires a well thought-out process work-flow and flexible database structure to store and manipulate the data. Moreover, internal policies and procedures must be created so that all those involved understand the project's purpose and goals, as well as the assigned roles of staff members and what is expected of them throughout the entire process. Once the data are published on-line and available to the end user, several steps must be taken to ensure that the customer will be satisfied not only with the data, but also the experience he or she will have in *obtaining* the data.

Archiving the Data Pile

The initial step of preparing and archiving the data is similar to taking an inventory of all items in a warehouse. Many organizations have dealt with this issue and have had to make "executive decisions" regarding unknown aspects of legacy data after project managers or veteran geologists retire or leave (Steinmetz et. al, 2002). Moreover, Steinmetz et al., (2002) pointed out that, within the petroleum industry, approximately 60 – 80% of a geoscientist's time is spent searching for data, while the remainder is spent organizing and analyzing the data. Ensuring that data files are properly cataloged and archived should therefore be a critical priority for any organization that provides data to the public. Documentation and ensuring data quality for legacy datasets are crucial in order to make the datasets meaningful and usable. Over the past several years, the DGGS has participated in the interagency Minerals Data and Information Rescue in Alaska (MDIRA) project. The MDIRA project has allowed the DGGS to overcome many challenges regarding its data archive, specifically database design (Freeman, 2001a, 2001b), restructuring and archiving legacy data by developing an appropriate geologic map model (Freeman and Sturmann, 2004), and writing and organizing metadata (Browne et al., 2003).

At the time of this writing, the total size of geologic geospatial data in the DGGS central server archive is approximately eight times greater than what it was in 2001, which is substantial, given that the average size of a single GIS layer file is on the order of tens to hundreds of kilobytes. There are also many large airborne geophysical datasets, personal databases, and datasets lingering on PCs that have yet to be archived. It is, therefore, the responsibility of project managers and authors to ensure that data files are not lost as tenured staff members leave the organization. In this day and age, the DGGS geologist is expected to solve complex geological problems in the field and in the office, gather information, process data, create a geologic map, and, additionally, archive and document all of the data associated with the project according to current division and FGDC standards. The time required to manage the magnitude of data associated with a given publication often dwarfs the time needed to analyze, understand,

and publish the data. Documentation of many geospatial datasets has been neglected because of geologists' need to initiate new mapping projects. Our hope at the DGGS is that the D3 Project will provide authors and geologists with an effective user interface that would allow them to manage and publish their data more efficiently, granting the geologist more time to "do geology."

What Should We Distribute?

The goal of the D3 Project is to provide datasets in formats that are easily adaptable to typical end-user systems. On the basis of the DGGS staff comments gathered during the project-planning phase, there will be two different file groups for data distribution (Table 1): digital data files and digital data product. After considering the two options, the DGGS decided to use "digital data files" in widely accepted data formats as the minimum standard for all forms of digital data distribution. Providing data to the DGGS's customers in the native data environment format is not the primary goal; however, it is an author-selected option for over-the-counter distribution (see below). Although providing all supporting native dataset environment files with the digital dataset may provide more information with end-user appeal, it is not the standard delivery we recommend. Key reasons for not choosing "digital data products" include: (1) the end-user will require the native software to use the data, (2) greater data liability (e.g. the misuse or misinterpretation of conclusions made by the DGGS), (3) high obsolescence risk, and (4) the need for special knowledge to distribute and use the data (e.g. data/software compatibility and end-user familiarity with the software). Examples of the digital data types distributed by the DGGS are shown in Table 2. These file formats include the most basic data format types that are capable of being adapted and used by a larger end-user audience. Because some data format types are technically proprietary formats (i.e. ESRI shape files, Microsoft Access databases, GeoSoft grid files), providing them in "more generic" formats would be unreasonable due to prohibitive file sizes or the complexity of the common format (i.e. distributing relational databases as ASCII, comma-delimited text files).

The proposed methods of data distribution include provisions for both over-the-counter and on-line distribution. The goal is to make each method distribute the same digital data, but in a different package. Custom distribution orders are always available.

Over-The-Counter Distribution

Digital data will be distributed "on-demand" on transportable media such as a CD-ROM for the over-the-counter (off-line) method. In this case, the publication number is the basis for distribution. Each CD-ROM will sell for $10 to cover the cost of the media, plus applicable postage/shipping costs. Over-the-counter contents could include:

- "ReadMe" file comprised of the table of contents, general information, and instructions in the use of the data or data product (standard)
- Metadata (HTML, ASCII text, and XML formats) (standard)
- PDFs of maps and reports for the publication (standard)
- Digital data files; format depends on data type as per Table 2 (DGGS Standard)
- Native dataset files, where different than digital data and if centrally stored and cleaned up (at author's option)
- Native dataset environment files, if centrally stored and cleaned up (at author's option).

Table 1. Comparison of two categories of data distribution.

Digital Data Files		**Digital Data Product**	
Pros:	Cons:	Pros:	Cons:
1. Simple to distribute 2. Wide audience 3. Easy to index 4. Consistent between projects and publications 5. Minimizes obsolescence 6. Smaller number of files 7. Not dependent on directory structure	1. User processing required before use 2. Annotation may be missing or in metadata 3. Requires export from native dataset 4. Larger file size	1. Data immediately usable/viewable 2. No file conversion (in native environment) 3. Full annotation 4. Contains all built-in logical relations 5. Often used by authors - formats already exist	1. Dependent on directory structure 2. Requires native data environment 3. Larger chance of data liability 4. High obsolescence risk 5. Requires more indexing 6. Difficult to manage 7. May require special knowledge to distribute and use
ASCII (comma or tab delimited), ESRI Shape, Geo-referenced TIFF, MSExcel, ArcExport		Digital data in native data environment (e.g., Geodatabase) *AND* supporting information like symbols, fonts, workspace files, base maps, etc.	

Table 2. Types of digital data formats.

Examples of digital data types	Digital Data Files (DGGS Standard)	Native Data Set Files	Native Data Set Environment Files
Tabular data	ASCII comma, tab delimited	Excel, Lotus 123, or other spreadsheets	NA
Vector data	ESRI shape files	ESRI files, geodatabase, MapInfo tab files	MapInfo workspace, ESRI Map document, fonts, symbol sets, shade sets, etc.
Raster data	TIFF and world file	TIFF and MapInfo tab files	
Grid data	ASCII comma or tab delimited, Geosoft grid or ESRI grid (size of ASCII files may be prohibitive)	ESRI grid files, MapInfo vertical files, ER Mapper grid files	
Relational databases	Native formats accepted here (i.e. MS Access), otherwise ASCII comma, tab delimited	Access, MySQL, or FileMakerPro database	Report, query or data entry documents (HTML, MSWord, Java, PSP, or ASP)

Furthermore, the data storage for the distribution files will use the existing directory structure (Freeman and Sturmann, 2004). All files distributed will be indexed in the database such that they can be located and copied onto distribution media on an “as needed” basis.

On-line Distribution

Digital data will be distributed on-line, free of charge, in compressed files to reduce volume and increase download speed. Compressed files will allow the DGGS to package metadata and other necessary documentation with the selected data as well as preserve any required internal file structure. Each compressed file will contain an individual digital dataset and metadata file, and will be listed with documentary information as an extension to the existing DGGS Publication Web Page for any given publication.

1. Each digital dataset distributed on the Web will display an abstract and have links to:
   - Compressed file containing a digital dataset as digital data files and metadata
   - Metadata file (including code set documentation) for the digital dataset
   - A link to the “ReadMe” File
   - Decompression instructions
2. Information about availability of over-the-counter “data on disk” will be included on the publication page with the following information:
   - Ordering instructions
   - A copy of the “ReadMe” file which includes the disk’s table of contents

The Data are Out There, Now What?

With the data files archived, indexed, and bundled into distinct datasets, and metadata written, it may be tempting to think that the job is done. At this point, however, certain aspects of the project are becoming relevant. For example, project managers and geologists must review the final layout of the publication page and datasets before they are officially posted to the Web, despite any previous quality assurance testing.

We are describing a major change in the functionality of the DGGS Web site. These changes will affect users and cooperators, which warrants some sort of notification. It would be beneficial to identify key end-user groups and notify them via the Web site itself, e-mail lists, monthly reports, meetings, or phone calls. Once end-users are aware of the new data-distribution service, it is imperative to provide effortless feedback methods with which these users can comment on data quality and ease of use, and submit suggestions. Similar to the open-source software community, the multitude of end-users are relied upon to find any remaining “bugs” in the system. Moreover, the DGGS will utilize database log files and web statistics to identify the most “popular” datasets and get a better understanding for what information is in demand.

LESSONS LEARNED

It was imperative when designing the D3 Project that the data distribution methods for DGGS staff were consistent and clearly stated. The D3 Project designers met with geologists and project leaders to discuss the distribution work-flow, user interface, responsibility assignments, and

details of particular types of datasets and archival strategies. The data distribution process should also be flexible to meet changing expectations and technical requirements of end-users. For example, breaking up the publications into several on- and off-line datasets provides flexibility and benefits those with small bandwidth or no Internet access.

Prior to distributing data to the public, an in-house inventory of existing data serves to identify which data are at risk. This process benefits both the distributor and the end-user by ensuring that the data adheres to current documentation standards, and by securing the data on more reliable media. Many agencies take the risk of storing and distributing data in proprietary data formats that may soon become obsolete or unreadable. With regard to such a risk, one has to ask, "Which, if any, software will be available 5, 10, or 20 years from now that can read the data?", and "When might the data become legacy data?"

In theory, data are always becoming legacy in status when software vendors upgrade their program packages, hardware becomes obsolete, and geologic maps are updated. Many agencies invest a large amount of time programming and creating scripts in the current software version, only to find that those scripts are worthless in the next program release. Similarly, storing precious data on only one type of archival media can be a terrible mistake. It is, therefore, up to project managers and authors to know when valuable data may be at risk and establish a legacy data recovery plan to prevent future data loss. Implementing a project such as this forces the agency to "clean house" and index valuable data.

Everyone involved with these kinds of projects must understand that documentation and data-quality information for every dataset are required. As a result, end-users will get consistent, quality data that are well-documented, which will allow them to have access to the information they need to use the dataset. If a user of a given dataset cannot find its documentation, he or she will more than likely (1) not use it, (2) attempt to use the dataset without proper guidance and understanding, or (3) use the dataset incorrectly or inappropriately. If project managers and authors take the time to document their data soon after it is created, the painstaking process of going back through tens or hundreds of datasets (some 20 years old), contacting retired staff members, and guessing about the details of a publication can be avoided. Moreover, by automating distribution methods to the greatest extent possible, the data can be delivered on demand. Since the freely provided data are already in digital form, easily searchable, well documented, and organized by dataset, users can focus on merging the data into their own projects and spend more time on analysis and understanding the implications of their scientific data and observations.

REFERENCES

Browne, C.L., Freeman, L.K., and Graham, G.R.C., 2003, The Alaska Division of Geological & Geophysical Survey's Metadata Policy Development and Implementation, in Soller, D.R., ed., Digital Mapping Techniques 2003—Workshop Proceedings: U.S. Geological Survey, Open-File Report 03–471, p. 201–208, available at http://pubs.usgs.gov/of/2003/of03-471/browne/index.html.

Davidson, Gail, 1998, Can we get there from here? Experiences of the Alaska Division of Geological and Geophysical Surveys, in Soller, D.R., ed., Digital Mapping Techniques 1998—Workshop Proceedings: U.S. Geological Survey, Open-File Report 98-487, p. 13–15, available at http://pubs.usgs.gov/of/of98-487/davidson.html.

DGGS Staff, 2005, Alaska GeoSurvey News: Alaska Division of Geological & Geophysical Surveys Newsletter 2005-2, 8 p., available at http://www.dggs.dnr.state.ak.us/pubs/pubs?reqtype=citation&ID=14595.

Freeman, L.K., 2001a, The DGGS Geologic Database: Putting Geologic Data Modeling into Practice: Alaska Division of Geological & Geophysical Surveys, Alaska GeoSurvey News, v. 5, no. 3, 3 p., available at http://www.dggs.dnr.state.ak.us/pubs/pubs?reqtype=citation&ID=14588.

Freeman, L.K., 2001b, A Case Study in Database Design: The Alaska Geologic Database, in Soller, D.R., ed., Digital Mapping Techniques 2001—Workshop Proceedings: U.S. Geological Survey Open-File Report 01-223, p. 31–34, available at http://pubs.usgs.gov/of/2001/of01-223/freeman.html.

Freeman, L.K., and Sturmann, A., 2004, Progress Towards an Agency-Wide Geologic Map Database at Alaska Division of Geological & Geophysical Surveys, in Soller, D.R., ed., Digital Mapping Techniques 2004—Workshop Proceedings: U.S. Geological Survey Open-File Report 04-1451, p. 9–14, available at http://pubs.usgs.gov/of/2004/1451/freeman/index.html.

Steinmetz, J.C., Hill, R.T., and Sowder, K.H., 2002, Digital Archives and Metadata as Mechanisms to Preserve Institutional Memory, in Soller, D.R., ed., Digital Mapping Techniques 2002—Workshop Proceedings: U.S. Geological Survey Open-File Report 02-370, p. 171–176, available at http://pubs.usgs.gov/of/2002/of02-370/steinmetz.html.

APPENDIX A

(Description of Terms)

Custom distribution: A custom distribution is a combination of data or data derivative that has not already been generated via the publication process. This may include requests for data reprojections, file format conversions, combining GIS layers from multiple projects or publications, statistical or spatial analyses, and excessively large amounts of data.

Dataset: A unique group of data that acts as a component of the publication. Examples include vector geologic features (i.e. bedrock, surficial, hazard polygons/lines/points), geochronology (i.e. spreadsheets, ASCII .csv), DEM data, electromagnetic anomalies, and grid data.

Digital data: Information that is ready for numeric or geographic manipulation with a minimum of conversion or preparation by the customer (e.g. Excel spreadsheets, formatted ASCII files, relational databases, geo-referenced raster files, geo-referenced vector graphics files).

Digital dataset: A logical, thematic, and geographic grouping of data, including any code sets (required to interpret the data). There may be one or more datasets per publication; a metadata document describes a digital dataset. Examples include GIS bedrock geology and spreadsheets that contain geochemical data related to a single publication.

Digital data file: Digital data in a file format that can be used across a wide variety of computing systems and meets the needs of most data consumers (See Table 1). These should be the standard formats that DGGS uses to distribute digital data.

Digital data product: Provides data and supporting information required to view the data in the native dataset environment (See Table 1). An example includes an ESRI Geodatabase and all supporting information like symbols, fonts, workspace files, base maps, etc.

Layer name: The name of the GIS layer, coverage, TAB file, or table name as defined in the metadata by the DGGS metadata extension, *Entity_and_Attribute_Layer_Name* (See Steps 4 and 5, Figure 1). If no layers exist in the metadata, the author may have to create layer names for their dataset within the application for the purpose of indexing their files.

Metadata: Metadata consist of information that characterizes data. Metadata are used to provide documentation for data products. In essence, metadata answer who, what, when, where, why, and how about every facet of the data that are being documented. Metadata written by the DGGS must conform to FGDC standards (http://www.fgdc.gov/metadata/geospatial-metadata-standards). Metadata will be distributed in three file formats to allow maximum readability and usability: Frequently Asked Question (FAQ) HTML, ASCII plain text, and Extensible Markup Language (XML).

Native dataset: Digital data in file formats that were produced by the software that was used to generate and process the digital data; the dataset does not include supporting native environment files (See Table 2). The user of these datasets may need access to the same software version that was used to produce the data.

Native dataset environment: The software operating system, hardware, and supporting files used by the producer to create, view, and process the dataset (See Table 2); it may be specific enough that it could be very difficult to replicate.

On-line distribution: Provides the e-mail and web browser customers with digital data in the form of compressed downloadable data.

Over-the-counter distribution: Provides the phone, mail, and walk-in traditional customers with digital data on some media (e.g., CD-ROM).

Project file: Any file found within the publication or project directory located in the DGGS directory structure on the central fileserver.