The Challenges and Benefits of Distributing Digital Data: Lessons Learned

Kenneth Papp [1], Susan Seitz [1], Larry Freeman [1], Carrie Browne [2]

Kenneth R. Papp
[1] Alaska Division of Geological & Geophysical Surveys
Geologic Communications
E-mail: kenneth_papp@dnr.state.ak.us
Phone: 907-451-5039

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

Presentation Outline

1) Project background and purpose

2) Problems to solve…questions that need answers

3) Data preparation

4) Process workflow

5) Benefits of distributing digital data

6) What we learned

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

“D3” Project Background and Purpose

- Primary goal of the Digital Data Distribution (D3) Project: Make DGGS geospatial data available to the widest possible audience
- Archive and index all digital project files and produce data set “packages”
- Provide authors with a graphical user interface to perform indexing and packaging tasks
- Packages will contain widely accepted “standard” file formats
- Utilize the World Wide Web to distribute on-line data sets and prepare off-line datasets

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

Problems to Solve

1) Cleaning up and archiving the mess of data…cracking the whip!

2) What should we distribute?

3) Metadata: the source of all information…and headaches!

4) Policy, procedure, and requirements

5) Designing a process workflow and flexible database structure

6) The data is out…now what?

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

Cleaning up and archiving the mess of data...cracking the whip!

- Many on-going projects at DGGS: focused on “cleaning up” legacy data
- Documentation of many geospatial data sets has been neglected because of the geologists’ need to initiate new mapping projects
- Review legacy data sets to document and upgrade the data to modern formats and documentation practices
- Documentation and ensuring data quality for the legacy data sets is key in making them meaningful and usable
- Need to make “executive decisions” regarding unknown aspects of legacy data after project managers retire/leave (Steinmetz et. al, 2002)

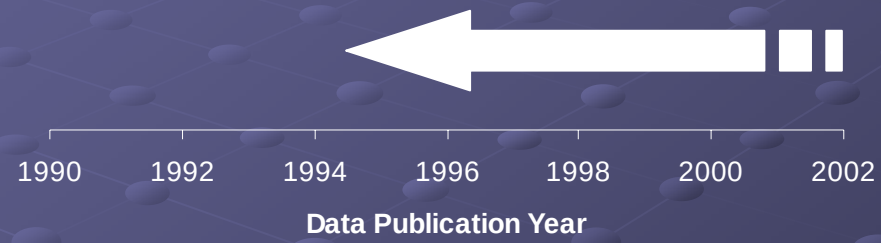
What should we distribute?

Examples: 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 or export files (E00)	ESRI coverage and 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 (size of ASCII files may be prohibitive)	ESRI Grid files	
Relational databases	Native formats accepted here (i.e. MS Access, MySQL), otherwise ASCII comma, tab delimited	NA	Report, query or data entry documents (HTML, MS Word, Java, PSP, or ASP documents)

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

Metadata: the source of all information… and headaches

- Legacy metadata project: 8th month of completion and greater than 60% of the legacy data has been recovered and its affiliated metadata has been written



- July, 2006: DGGS will test and implement the Metavist metadata writing tool (D. Rugg, USDA Forest Service) to write FGDC-compliant metadata

- Metadata saved as XML documents and loaded into the Oracle database using Oracle XML DB functionality

Policy, procedure, and requirements

- Uniform data distribution procedure that complies with Alaska statutory requirements
- Clear digital distribution methods for DGGS staff to consistently use
- Flexibility to meet changing expectations and technical requirements of end-users
- Understood that documentation and data-quality information for the data sets is required
- Automate distribution methods to the greatest extent possible so that data can be delivered on demand
- Incorporate feedback from geologists and end-users >>> cannot please everyone

Designing a process workflow and flexible database structure

- Chicken and the egg: database structure or process workflow?

STEP 1: DISTRIBUTION PREPARATION
Metadata, cleaning up project files, archiving data on network

STEP 2: FIND DATASET
Entry point to internal application.
Log in, find dataset via publication information

STEP 3: FIND PROJECT FILES
Browse to archive location via application, locate files for distribution

STEP 4: INDENTIFY LAYERS
Enter and describe layer names of the dataset

STEP 5: INDEX FILES BY LAYER
Associate files to be distributed with dataset layers

STEP 6: CREATE DISTRIBUTION PACKAGE
Identify files to be distributed together

STEP 7: REVIEW DISTRIBUTION PACKAGE
GIS Manager reviews distribution packages for data quality

STEP 8: PUBLISH DISTRIBUTION PACKAGE
Final approval and release to the public

Note: Author tasks outlined in red

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

The data is out there…now what?

- “D3” project completion by the end of 2006
- Project managers/geologists must review the final layout and data after publication to the web
- Getting the word out: identify key end-user groups and notify them (i.e. web site, email lists, monthly reports, meetings, phone calls)
- Provide feedback methods for end-users regarding data quality, ease of use, and future suggestions
- Utilize database log files and web statistics to identify most “popular” datasets

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

Benefits of distributing digital data

- Forces you to “clean house” and index the data
- To all geologists: “If you want your data on-line, you have to write metadata.”

  Result: End-users will get consistent, quality data that is well documented
- Breaking up the project into several on- and off-line data sets provides flexibility and benefits those with small bandwidth or no Internet access
- The available “raw” data can be quickly implemented into other projects and used appropriately (i.e. documentation)
- Project managers, geologists, and data managers better understand the “entire” publications process

What we learned…

- It is worth using resources to “clean up” legacy data sets
- Not enforcing a consistent archival structure for project data is not good
- You can’t make everyone happy. An even balance between regulations, consistency, and user freedom is always difficult.
- Failed contracts: know when to pull the plug
- Database managers and programmers can benefit by thinking like a geologist…and vice versa.
- Decrease the whaling and gnashing of teeth…document your data as you create it!

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us

Thank you!

Kenneth Papp, Geologist
kenneth_papp@dnr.state.ak.us