Alaska Division of Geological & Geophysical Surveys

Seven years and two terabytes:

Our experiences in paper to digital document conversion.

http://www.dggs.dnr.state.ak.us/

Overview

- DGGS scanning projects - history and scope
- Workflow and process steps
- In-house scanning vs. outsourcing
- Equipment selection and cost
- File management
- Archive and distribution formats
- Ongoing maintenance

DGGS scanning projects - history and scope: **Justification**

- Relative to many other states, Alaska is significantly burdened by high costs of geologic data acquisition.
- In many regions of the state our geologic knowledge is limited to pioneering work done by earlier generations of scientists.
- Until recently, convenient and affordable access to comprehensive geologic datasets wasn't available to folks outside of Anchorage, Juneau or Fairbanks

DGGS scanning projects - history and scope: **History**

- 2000 – DGGS received funds from a multi-year, multimillion $$ - USGS data preservation grant
- Among other related tasks, the grant obligations entailed two imaging components:
  - Scanning DGGS and Alaskan topic USGS geologic reports
  - Cataloging and archiving (physically and digitally) our library of pre-publication and unpublished legacy data.

DGGS scanning projects - history and scope: **Current Status**

- DGGS completed the publication scanning project in three chunks entailing about two years of work for each chunk.
- Our digital image library contains:

  - 15,000 citations

  - 27,000 files/image groups
- I am a geologist with a work experience background in library science and web design.
- I’ve worked for DGGS since 2000 and inherited the publications scanning project in 2002.
- More recently, I’ve also inherited our legacy data archiving project.

Workflow and process steps:

- Develop inventory and cataloging methodology
- Develop storage and distribution methodology
- Define image specifications
- Quality control methodology
- Equipment purchases and maintenance
- Plan for ongoing maintenance of dataset
- Develop a planned workflow that uses time and money most efficiently
- Personnel needs – cataloger, manager, scanner, programmer, quality control
- Usage tracking

In-house scanning vs. outsourcing: **DGGS Experiences**

- DGGS has utilized both in-house and outsourced scanning services
- During the initial years of our project (2000-2002) outsourced scanning of large format maps simply wasn't available and approximately 4,000 oversize sheets were scanned in-house
- Subsequent drops in technology/scanning prices and rises in State personnel costs have rendered in-house scanning of large document batches impractical for us
- We have purchased and continue to maintain a mid-volume sheet-feed and a wide format scanner for delicate items and small volume batches

In-house scanning vs. outsourcing:

Outsourcing

- In addition to cost effectiveness of scanning itself DGGS has found outsourcing to be far more effective in a number of other ways:
  - Greater control of project timing
  - Stronger recourse for quality control issues
  - Less responsibility for equipment purchasing and maintenance

In-house scanning vs. outsourcing:

Contract Specifications

- I've had the opportunity to work with three different scanning contracts ($30,000-$60,000), two of which I've written myself
- This is what I've learned:
  - Know your inventory, test the process, understand the resultant product
  - Define your content: page count, document size, image type, media type, document condition, etc…
  - Define how the submittal documents will be organized and how resultant files will be named, organized and delivered
  - Define your scanning specifications and understand why you've selected those definitions

In-house scanning vs. outsourcing: Contract Specifications (cont.)

- Define your quality control procedure and the consequences of poor scan quality or incompliance with contract terms
- Take advantage of the “Request for Proposal” process such that you can assign the contract based on experience, references, and qualifications
- Require a test prior to contract approval that verifies their attention to detail and understanding of your documents and imaging requirements

Equipment selection and cost: **Scanners**

- DGGS is on it's third wide format scanner in seven years ($15,000 - $20,000 each)
- Recent drops in pricing have allowed us to also purchase a mid-volume business class, 11x17 flat bed/sheet feed combo scanner ($6,000)
- Maintenance costs and extended warrantee pricing on scanning equipment seems to be currently running at about $1,500 per year or incident
- Moral of the story: this stuff ain't cheap! You probably can't afford to own unless you can also afford to replace it and/or maintain it.

Equipment selection and cost: **File Storage**

- Until about 2003 our primary file archive media was CD (the best affordable technology at the time)
- To protect ourselves against data loss we stored and generated 3 duplicates of each CD (about 700 CDs/set)
- It was a complete and total nightmare to quality control and maintain
- Modern, cheap server hard-drive storage with tape back-up makes it a pretty obvious choice for 2007 and the near future,
- Lower budget projects w/o ready access to server drive storage w/tape back-up may wish to investigate interagency availability before resorting to PC drive storage and/or CD

File management

- Schedule back-ups, test your back-ups!
- Eliminate file duplication to the fullest extent possible
- Name files and directories (folders) in a predictable and uniform manner, this facilitates manipulating the files programmatically
- Develop tools (scripts, programs) for reconciling file names in your catalog with the files on your drive
- Plan for file migration…

Archive and distribution formats

- DGGS currently uses 300dpi single page uncompressed TIFF as our archive format,
  - Our TIFFs are commonly utilized in GIS and drawing applications and 300 dpi adequately captures an appropriate level of detail while generating a usable file-size,
  - Finer resolutions would pick up interference information such as smudges on the paper w/o significant improvements in actual data capture
- Adobe Acrobat 7 (“hidden image compact” OCR setting) as our web distribution format
- We’ve tried and abandoned a few other formats for archiving and distribution, these include:
  - Lizardtech MrSid
  - Adobe Acrobat “formatted text and graphics” OCR setting
  - Multi-page TIFF
  - JPEG compressed TIFF
  - LZW and group 4 compressed TIFF
  - Higher and lower TIFF resolutions
  - Geo-referenced images, TIFF and MrSID
  - PDF files generated by Image Magik