Agency-wide Implications for Alaska DGGS building a Statewide Geologic 1:100,000 Compilation from an NCGMP09 Multi-map Database

By Jennifer Athey and Michael Hendricks
Alaska Division of Geological & Geophysical Surveys
3354 College Road
Fairbanks, Alaska 99709

Jennifer Athey
Telephone: (907) 451-5028
email: jen.athey@alaska.gov

Michael Hendricks
Telephone: (907) 451-5029
email: michael.hendricks@alaska.gov

Abstract

The Alaska Division of Geological & Geophysical Surveys (DGGS) has undertaken the challenge of developing a multi-map, multi-user database model based on the single-map NCGMP09 geologic map schema developed by the USGS and state geological surveys. The new multi-map model is intended for national use, as is a pilot data-sharing protocol to be developed with the model. Over the next three years, DGGS is seeking interested individuals to take part in discussions via teleconferences to provide input on the needs of geologic surveys and other organizations and help develop the specifications of the database model and data-sharing protocol. A multi-map geodatabase will help DGGS meet the future goal of a 1:100,000-scale Alaska compilation, and provide a vehicle for other geologic surveys and agencies to organize and share their own geologic data.

Slide 1. DGGS has planned to incorporate a statewide "enterprise" geologic database into its business processes for almost two decades. At the end of 2000, we began fleshing out a schema based on the North American Data Model (NADM; Freeman, 2001). By 2005, DGGS had created a database model that encompassed most of the functions of the survey, including field, geologic, analytical, and archival data; publications; and some administrative tasks like invoicing. Population of most of the database at that time, however, was delayed, because the amount of overhead needed to run Oracle Spatial and Esri's SDE, a separate Esri product at the time, was beyond the resources of the survey. Ten years later we still recognize the importance of a comprehensive data management system, and we are re-engaged with a new plan, new staff, and the new geologic data model NCGMP09 from which to build. Currently in DGGS's enterprise database environment, publications and physical materials are successfully managed, field and analytical data are partially managed, and geologic data are largely unmanaged. With the goal of a seamless 1:100,000-scale map of Alaska in the future, DGGS is beginning a project to develop and implement a database that can contain multiple geologic maps, allow for multiple users at the same time, and provide standardized across the various data sets.

Slide 2. DGGS would like to go about this project in a different way—by engaging the geologic community—for two reasons. First, the Environmental Protection Agency's (EPA) Environmental Information Exchange Network Grant Program (https://www.epa.gov/exchangenetwork) has tentatively funded the creation of a multi-map geologic database and data-sharing schema over three years. The Exchange Network's philosophy is to create communication and data-sharing networks among data users and producers in the EPA, states, and other agencies, with the goals of more robust products and better decision making to support environmental and health issues. The EPA names the geologic community as a desired partner to acquire A-16 geospatial, themed data (https://www.whitehouse.gov/omb/circulars_a016_rev) and update national data sets such as the National Geologic Map Database (http://ngmdb.usgs.gov/). Essentially, the Exchange Network encourages national data standardization and dissemination efforts rather than smaller, regional work. It is DGGS's hope that, by opening up the development process to the thoughts and needs of other geologists and their surveys, we will be able to create a schema, workflow, and documentation that will benefit other organizations as well as our own, and move forward the ability for everyone to efficiently exchange geologic information.

The second reason DGGS would like to engage the geologic community in this project is simply that the products will benefit from the collective wisdom and points of view of a variety of people. Over the years of attending Digital Mapping Techniques (DMT) workshops, we have noticed that geological surveys conduct business in similar ways and have similar data management needs. Starting this project with information from state surveys and other geology-related agencies who already have experience with multi-map geodatabases, or who might be interested in this technology in the future, will help ensure that the products are applicable to a larger audience and available more quickly.

Anyone is welcome to join the discussions. Please contact Jen Athey at jen.athey@alaska.gov for more information. Alaska DGGS will facilitate monthly or bi-monthly teleconferences among interested geologic community members. Ideas from the community will be followed with testing, and documented results will be made available to the discussion group by DGGS. We anticipate this will be an iterative process. Progress and results will also be presented at future DMT workshops.

Slide 3. DGGS's EPA-funded project includes two additional goals besides development of a geologic database model and data-sharing protocol (goal 2). The complete project includes the development and population of a radon database for Alaska and radon data-sharing schema based on work by the Centers for Disease Control and Prevention (CDC) Environmental Health Tracking Network and other states (goal 1), and the implementation of a web mapping application that portrays both geology and radon data (goal 3). Completion of goals 1 and 3 will not involve geologic community discussion.

The development of data-sharing protocols are an important component of the EPA's Exchange Network data-sharing philosophy. As defined in the proposal, the data-sharing protocols for radon and geology will be well documented, predictable, and easily ingestible by GIS software. DGGS proposed to create a "pilot" data-sharing protocol for geology. Because standards development is a long process and requires buy-in from potential users, a version of the geology data-sharing protocol will not likely be finalized and accepted nationally by the end of the three-year project.

Slide 4. Although creating a 1:100,000-scale compilation map of Alaska is a DGGS goal, sparse data are currently available and correctly formatted to be entered into a database in support of this effort. DGGS has six maps in the NCGMP09 format, but the maps are not fully NCGMP09 compliant and lack tables such as the Description of Map Units and Glossary. Consequently, the best reason for initiating the construction of the database at this time is for the ease of future standardization and data management. The longer we wait to create a system to manage our geologic map data, the more data we will have that is not compliant with the database.

DGGS wonders how many geological surveys already have a database to manage their geologic data. Of the state surveys that don't have a geologic database, how many might be interested in developing one if the needed resources were relatively minimal?

Slide 5. The USGS's NCGMP09 geologic data model is designed to store the geologic data (polygons, lines, and points); some metadata; and the symbology needed to create a single, traditional geologic map (http://ngmdb.usgs.gov/Info/standards/NCGMP09/). In its current single-map version, the data model is a powerful but flexible tool capable of standardizing the geologic information of an organization, one publication at a time. There are advantages, however, to building a version of the data model that can hold multiple maps.

1. Because the data model is designed to be flexible, data can be stored appropriately in multiple ways, such as saving the map's "neatline" (map boundary) in either the "OtherPolys" or "ContactsAndFaults" table. With different personnel entering data and changes in thought over time, an organization's geologic data might ultimately become only partially standardized. Keeping map data together in one location will enforce more rigorous standardization and improve accessibility.
2. A multi-map database would facilitate web-accessible, aggregated geologic data and creation of data products aided by standardization, such as compilations.
3. Information in the NCGMP09 Glossary, DataSources, DescriptionOfMapUnits, GeologicEvents, and StandardLithology tables can be used to describe data in multiple maps. In the single-map NCGMP09 schema, these tables must be repeated for each publication.
4. Beyond the traditional, published geologic map, the data model is flexible enough to hold unpublished data, atypical data releases, field interpretations, and analyses, if desired.
5. Making the database editable by multiple users can further increase efficiency.

A more thorough discussion of the benefits and detriments of a multi-map, multi-user database follows on slides 9 and 10.

Slide 6. DGGS's IT strategy and general project specifications indicate development in open-source PostgreSQL database platform with spatial data access, administrative control, and multi-user functionality from ArcGIS for Server and its built-in Spatial Database Engine (SDE) component. ArcGIS for Server/SDE is currently the only software available that provides all of the functionality needed for this project. The single-map NCGMP09 standard specifies the use of the proprietary Esri geodatabase format, because of the pervasive use of ArcGIS in the USGS and state surveys. (The standard also requires the open file format *shapefile* as an end product.) Further, we know of two multi-map, multi-user geodatabases based on NCGMP09 in other organizations that are or were running PostgreSQL and ArcGIS for Server/SDE. We hope to utilize the knowledge and efforts of these organizations.

In 2015–2016, the NCGMP09 workgroup, headed by the NGMDB project and composed of interested individuals from state and federal geologic agencies, discussed two basic structures for a multi-map version of the schema. In scenario 1, information from various maps would be aggregated into tables of similar data that follow the NCGMP09 standard. For instance, all contact and fault line data would be aggregated into the table ContactsAndFaults, and a primary key would indicate the map from which the data were harvested. To our knowledge, scenario 1 has not yet been attempted, although an Alaska-commissioned analysis of the NCGMP09 standard by Esri in 2013 suggested it is a viable option.

In scenario 2, individual maps would be stored in separate sub-databases, where tables of like data would have the same NCGMP09 data structure. Data types could be compiled into aggregated layer files as needed; however, using compiled layer files would also require that records in sub-databases have a primary key referencing the map data source. In scenario 2, some tables such as the Glossary and DataSources could be referenced by all of the map sub-databases. Each scenario has benefits and detriments, and must be evaluated against the key considerations of the data management system listed in slide 6.

Slide 7. For our own purposes, DGGS is considering developing a database model that contains a modified version of NCGMP09 as well as historical, unpublished, field, and analytical data. Only the geologic map portion of the schema is a grant deliverable. This slide depicts the relationship between historical data and data that are actively being worked. Like many geological surveys, DGGS has an abundance of rasterized geologic maps. In the future, only the most important maps may be digitized, attributed, and added to the database to help with the 1:100,000-scale layer. Since this is a relatively time-intensive process, many raster maps will never be digitized and attributed; instead, they may be used as background images and feed web map services.

Note that this database model includes compilation data sets at 1:100,000 and 1:500,000. A senior geologist will be assigned to vet and add incoming data to the 1:100,000-scale compilation. As future mapping is conducted in the state, care will be given to edge match existing data and correlate with known geologic units as appropriate. DGGS will essentially build the 1:100,000-scale compilation as we map new areas.

Slide 8. The second part of the model shows the relationship of field and analytical data to geologic maps. Because geologic maps are synthesized from a variety of data sets, there are many different data types to be preserved in the database. Not all of the data may be portrayed on the map, even though they may have influenced the geologic interpretation. In addition, ephemeral interim products may be important to save. Reproducibility is valuable—one can argue that the process we use to reach an interpretation is just as important as the interpretation itself.

DGGS questions whether analytical and field data should be housed in PostgreSQL outside of the main geodatabase and ArcGIS for Server environment. In most cases, field data and analyses are tied to a field station or sample point, and that location is recorded in the NCGMP09 Stations table. In PostgreSQL, analytical values could be queried into a database "view", or virtual flat table, and the view could be a table in the SDE geodatabase. The analytical data could then be joined with locations in GIS and synthesized.

Slide 9. Because "multi-map" and "multi-user" environments satisfy two very different needs, implications for each are described separately.

The most important difference between the current NCGMP09 model and the "enterprise" geodatabase is its ability to hold multiple maps. The USGS recognized that the NCGMP09 standard is a springboard to the development of multi-map databases for mapping agencies, and a step toward the creation of a national archive of geologic map information under the National Geologic Map Database project (NCGMP, 2010). The benefits of this change in the schema are related to an increase in standardization and accessibility. In particular, standardization has a positive long-term effect on business processes, products, data quality, and services. Expanding the NCGMP09 schema will also change its scope from one of representing a single, traditional geologic map to managing the geologic data of an organization, which comes with a host of issues such as user permissions, edit tracking, and managing vocabulary and domains. In the short term, resources will need to be spent on creating, testing, and implementing new business processes in the organization, i.e., short-term pain for long-term gain.

Slide 10. The obvious benefit of a geodatabase that can be edited by multiple people at the same time is increased efficiency. This database specification will be of greater interest to larger organizations and collaborative mapping teams. Some organizations may choose not to implement multi-user capability, because either their workflows don't demand it or their resources are too limited. If multiple people are editing a data set, fairly time-consuming data management procedures must be strictly followed to maintain data cohesion.

Because DGGS has never had multi-user capability in the GIS environment, our current workflows don't necessitate it. While most of our work will not involve multiple people editing and accessing the same data, multi-user capability will be useful for some tasks. We anticipate that more collaborative work will occur digitally in the future. Examples of multi-user work include:

- Multiple geologists mapping in adjacent areas, with ongoing compilation as the work progresses;
- Multiple GIS analysts or geologists updating the same data set;
- Editing and final map display decision making between geologist and cartographer; and
- Creation of an overarching compilation while synthesis and cartography continues on individual data sets.

Slide 11. To maximize the benefits and minimize the detriments of such a geologic model, DGGS has short- and long-term goals. In the near future, our plan concentrates on implementing the geodatabase without derailing current map production. We will provide training and make incremental changes so staff are not overwhelmed with new procedures. Obtaining buy-in from staff will be critical to the success of the project.

Far-future goals are to take advantage of the benefits of standardization and re-invest in DGGS's map making process to make it more efficient. Updating a 1:100,000-scale statewide compilation and system maintenance will be operational tasks in perpetuity. A first version of the 1:100,000-scale compilation might not be completed for 70 to hundreds of years from now, depending on funding.

Slide 12. Data cannot be used unless they can be shared. In the life cycle of data, data usage is the natural successor to synthesized data in a database (Chisholm, 2015). We often think of geologic data sharing in terms of the world's needs and safety, such as energy resources, climate change, and earthquake mitigation, but there are many uses for geologic data in each of the federal departments listed on the slide. The EPA is just one of many government agencies needing better access to geologic data. Availability of a data-sharing protocol, the other grant deliverable, is an important objective of the EPA's Exchange Network, because the EPA recognizes that understanding geologic information will help them make better decisions about the environment. It is critical that the geologic community provide data that are easy for others to ingest, both in terms of file format and type of information, because the flow of understandable geologic data ensures that our work is relevant to other organizations and the public. Work that is not relevant will not get funded.

Slide 13. Sharing a discrete geodatabase or group of shapefiles works reasonably well for just one map, but sharing data from multiple maps involves the transfer of a large amount of predictably formatted information. A data-sharing protocol for multi-map geologic data is required. The only existing standard that might meet the community's needs is GeoSciML, a markup language for geologic data, which has multi-national support (http://www.geosciml.org/). An abbreviated version of the standard called GeoSciML-Portrayal is developed for use with web services. The standard is not yet widely accepted by the geologic community and will be examined during this project. Ideally

DGGS would appreciate an easy-to-format protocol that can be translated from the geodatabase, perhaps using the web service tools that already exist in ArcGIS for Server.

The grant deliverable will be a "pilot" data-sharing protocol, since it is unreasonable to expect that the protocol selected through this project will be finalized and accepted nationally in the short timeframe of the grant.

Slide 14. Developing a national multi-map geologic database model and "pilot" data-sharing protocol is an ambitious but doable multi-year project, given the extensive work already completed on the NCGMP09 schema. DGGS hopes to collaborate with the geologic community through teleconferences and DMT meetings to determine the specifications and structure of the multi-map geodatabase and data-sharing protocol. In July, DGGS will send out a notice over the DMT listserv asking for interested people to join a teleworking group to discuss the "enterprise" database. DGGS will test the discussed scenarios and report findings back to the workgroup and community. By providing updates and feedback during the development process and documentation and coding for the final deliverables, DGGS expects that the products produced through the grant will be eminently useful.

Slide 15. Quilts are a great analogy for many processes in geology such as building compilations, conducting mapping, and finding repeating patterns. However, in this case, we submit the quilt-making process as a compelling analogy for teamwork. This particular quilt is called a scrap quilt, and it is made from the dregs of a dozen people's fabric chests. The fabric scraps are leftover from baby, wedding, and retirement quilts that chronicle the lives of DGGS's staff. In the analogy, each piece of fabric is a person's experience and expertise. Like quilters combining fabrics together into a deliberate but unique pattern, group discussion pieces our experiences together into actionable ideas. By binding those ideas together, we can build something useful!

REFERENCES

Chisholm, Malcolm, 2015, 7 Phases of a data life cycle [website]: Information-Management.com, accessed 6/16/2016. http://www.information-management.com/news/data-management/Data-Life-Cycle-Defined-10027232-1.html

Freeman, Larry, 2001, 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. http://pubs.usgs.gov/of/2001/of01-223/freeman.html

National Cooperative Geologic Mapping Program (NCGMP), USGS, 2010, NCGMP09—Draft Standard Format for Digital Publication of Geologic Maps, Version 1.1, *in* Soller, D.R., Digital Mapping Techniques '09—Workshop Proceedings: U.S. Geological Survey Open-File Report 2010-1335. http://pubs.usgs.gov/of/2010/1335/pdf/usgs_of2010-1335_NCGMP09.pdf

Agency-wide implications for Alaska DGGS building a statewide geologic 1:100,000 compilation from an **NCGMP09 multi-map database**

JENNIFER ATHEY AND MICHAEL HENDRICKS

ALASKA DNR/DIVISION OF GEOLOGICAL & GEOPHYSICAL SURVEYS (DGGS)

DIGITAL MAPPING TECHNIQUES 2016, TALLAHASSEE, FLORIDA, MAY 22-25
PRESENTED BY JENNIFER ATHEY, MAY 25, 2016

Forward-looking statement

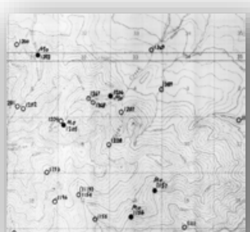
DGGS is planning to develop a multi-map, multi-user version of NCGMP09 and pilot data-sharing schema. We hope to harness the collective experience and knowledge of the geologic community to create well-documented products that can be replicated by other agencies. A data-sharing mechanism or service, particularly if it includes non-specialized language, will increase the usability of our data by other geologists, scientists, and the public.

Specifications to be determined.

Overview of EPA project

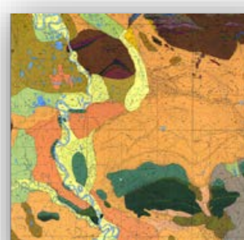
Goal 1

Develop radon database for Alaska and data-sharing schema



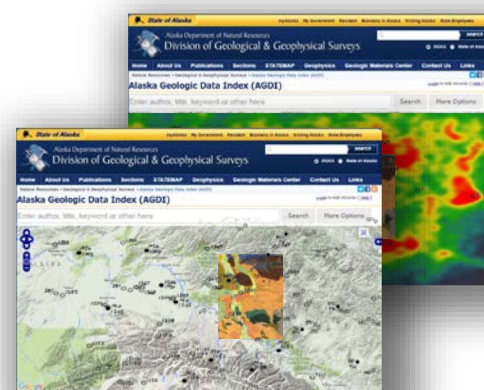
Goal 2

Develop "enterprise" version of NCGMP09 database and data-sharing protocol



Goal 3

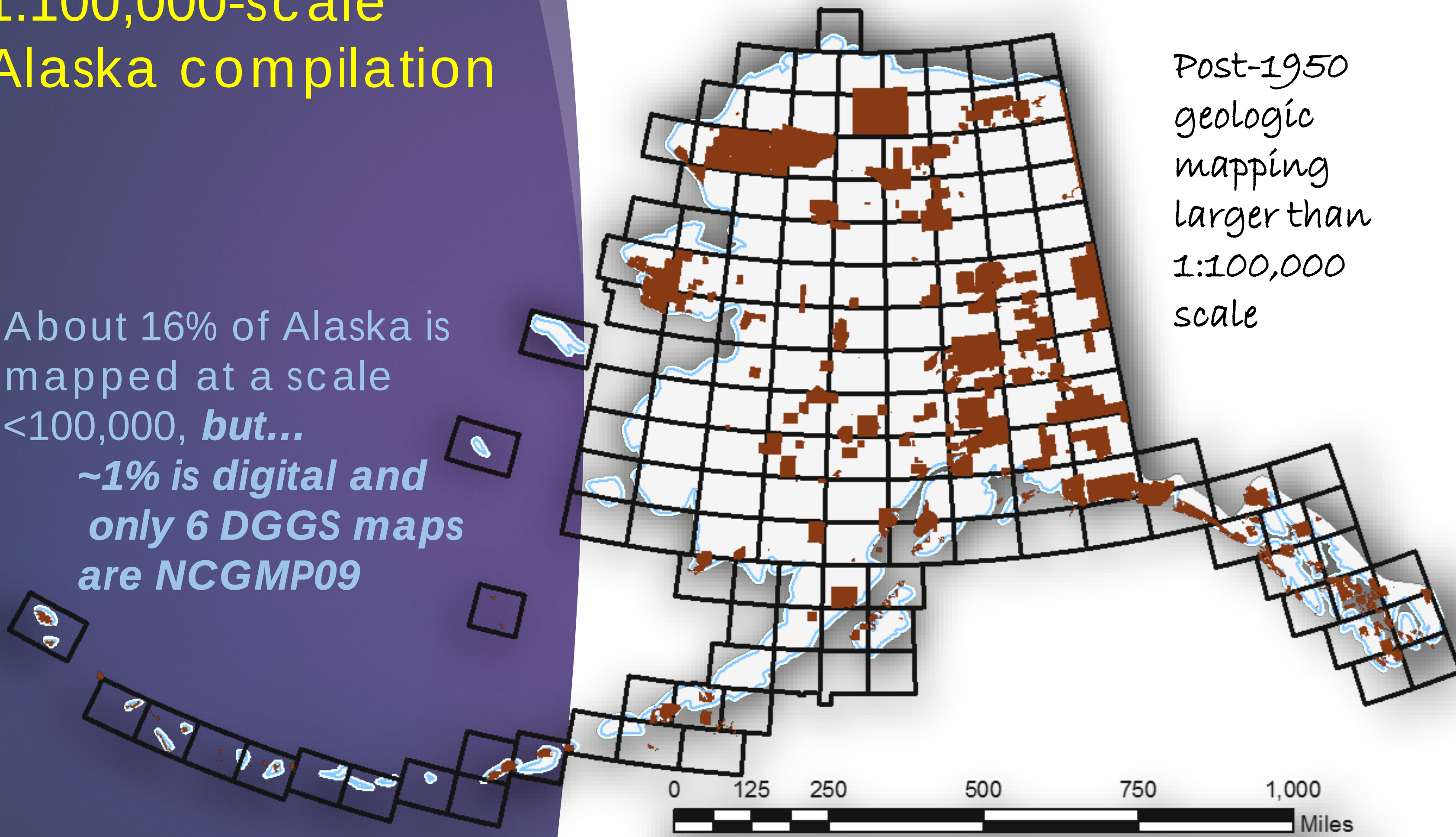
Create predictive geology-radon web map with radon "heat" map overlay



DGGS goal to build 1:100,000-scale Alaska compilation

About 16% of Alaska is mapped at a scale <100,000, ***but…***

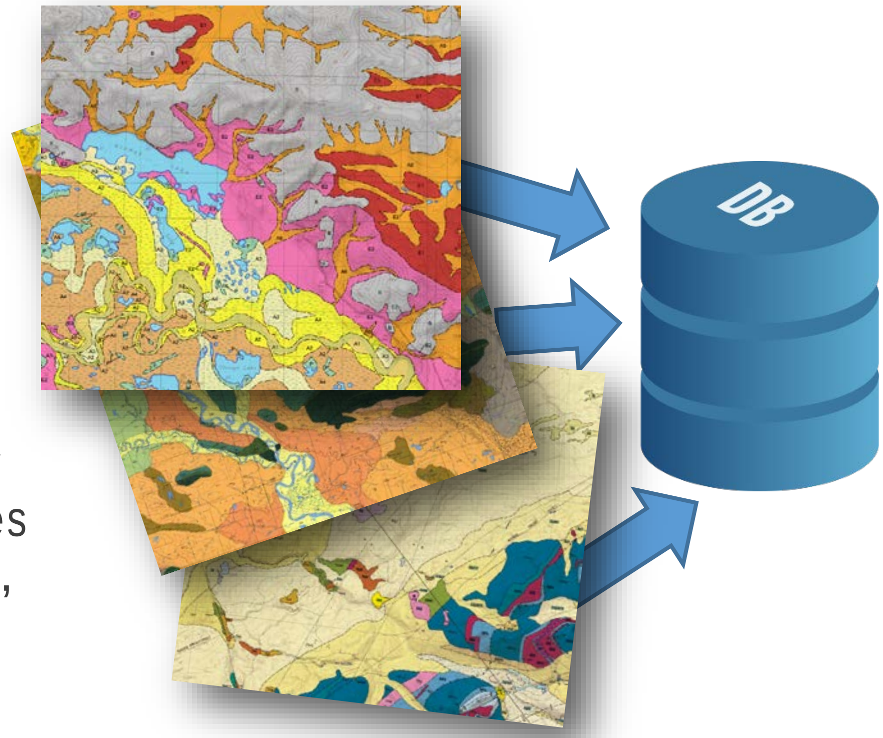
~1% is digital and only 6 DGGS maps are NCGMP09



Post-1950 geologic mapping larger than 1:100,000 scale

What is a multi-map, multi-user NCGMP09 database?

- A controlled container for agency-wide spatial data (not just geologic maps?)
- A vehicle to standardize geologic data, increasing accessibility and enabling digital products
- A way to increase efficiency through standard procedures for map production, analysis, compilation, and archiving



Setup and considerations

Technical

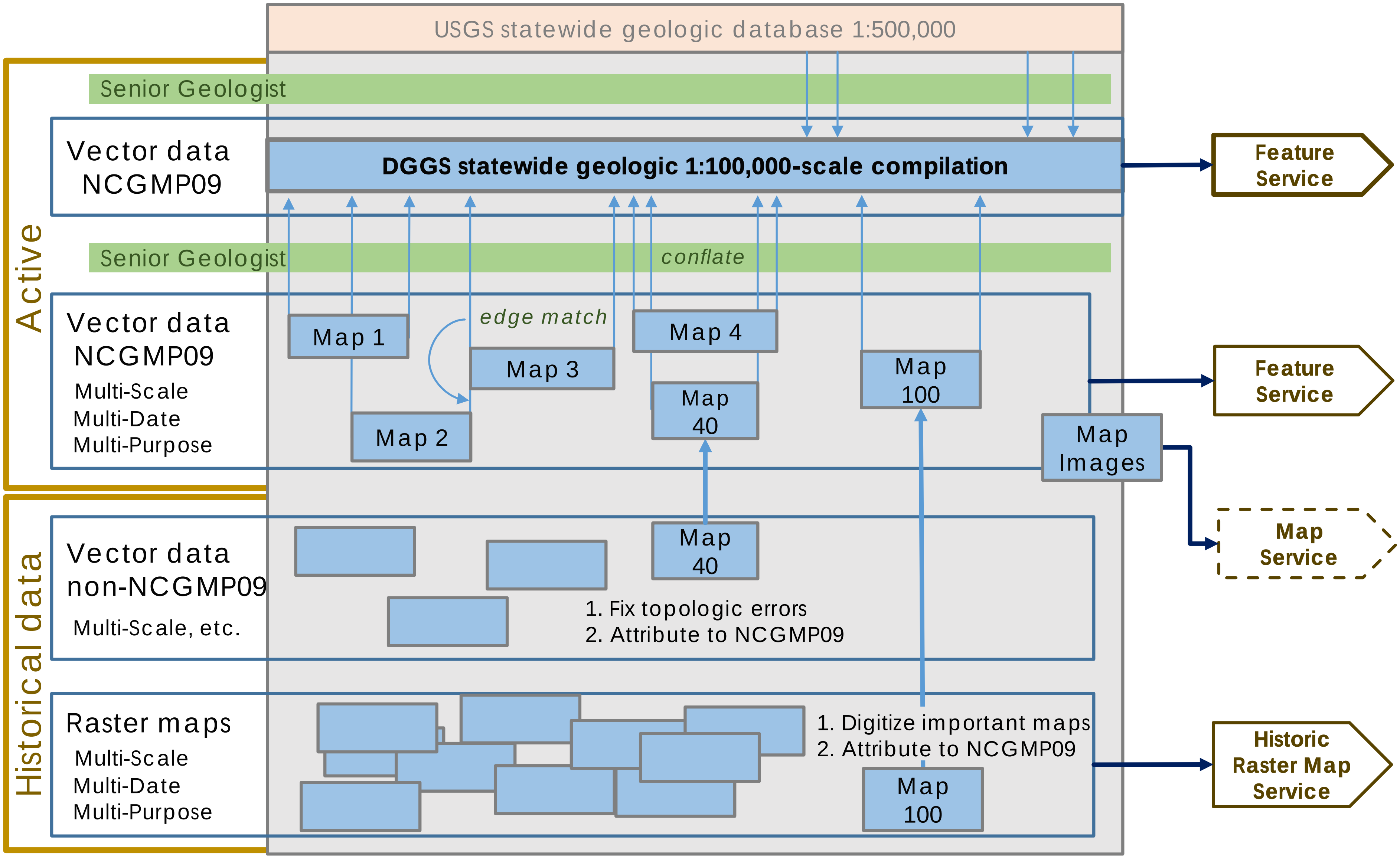
- PostgreSQL
- ArcGIS for Server/SDE

Geodatabase structure?

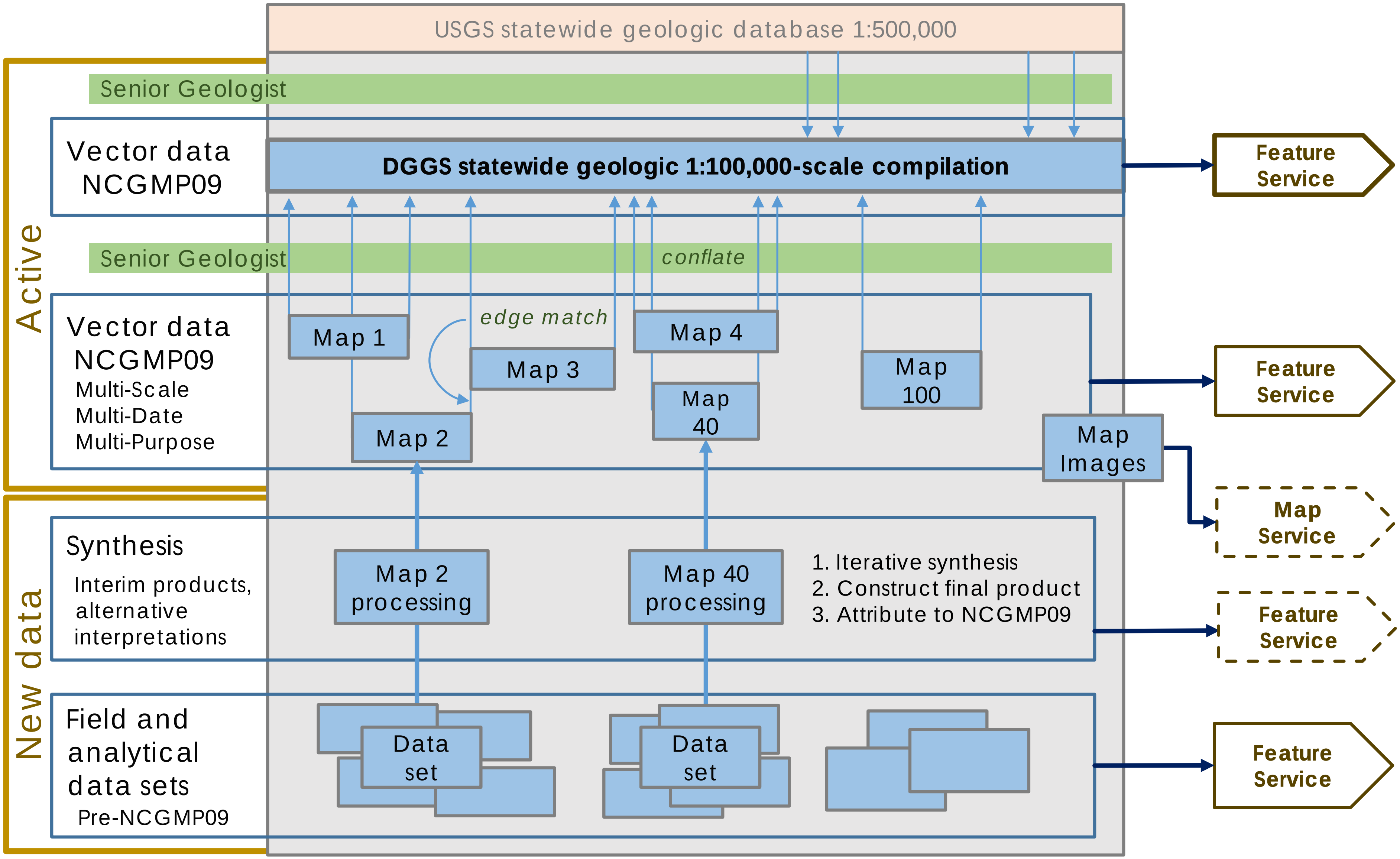
- Maps combined into single tables with a primary key
- Maps in sub-databases, data are aggregated into layer files
- *Others?*

- Reasonable speed of access to data (draw time)
- Multi-scale, multi-temporal data sets
- Scalable – some data sets will be huge
- Queryable across multiple maps
- Allow use of production tools
- Allow single and multi-map unit descriptions
- Ease of use for staff
- Multi-user access to data
- Integration with data in other databases
- *Others?*

Draft DGGS database model, part 1



Draft DGGS database model, part 2



Multi-map implications

Benefits

From standardized, aggregated data…

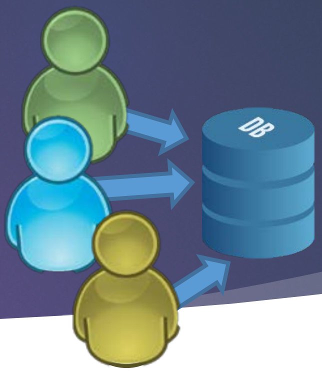
- *Greater efficiency* from production tools
- *Better products* from expected values, more robust data sources
- *New products* such as compilations, derivatives, web services, outreach

Detriments

From setup of a new system…

- *Time sinks* from learning new processes, synthesizing more data, converting data to NCGMP09
- *Overhead* from database administration and QC of new data
- *Untried processes* like converting non-traditional maps and data to NCGMP09

Multi-user implications



Benefits ➕

From multiple people working on the same data…

- *Greater efficiency* from increased production
- *Better products* from multiple people interpreting data, and using different analyses and data sources

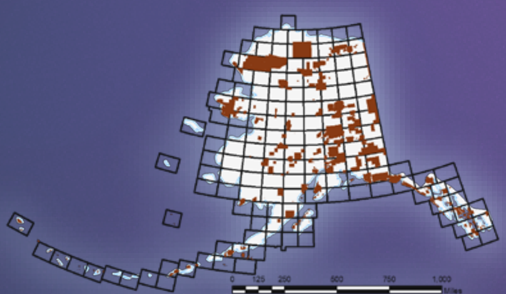
Detriments ➖

From setup of a new system…

- *Time sinks* from learning new processes and resolving GIS data conflicts
- *Overhead* from system and database setup and maintenance
- *Untried processes* such as creating a digital compilation in the field

Planning for the future

- Cover the state in 1:100,000- or larger-scale mapping
- Create a 1:100,000-scale compilation



70 years? 400 years?

Short term

- Organizational buy-in (in progress)
- Hire staff (done)
- Develop infrastructure (in progress)
- Database design (in progress)
- Develop database and admin settings
- Testing
- Staff training

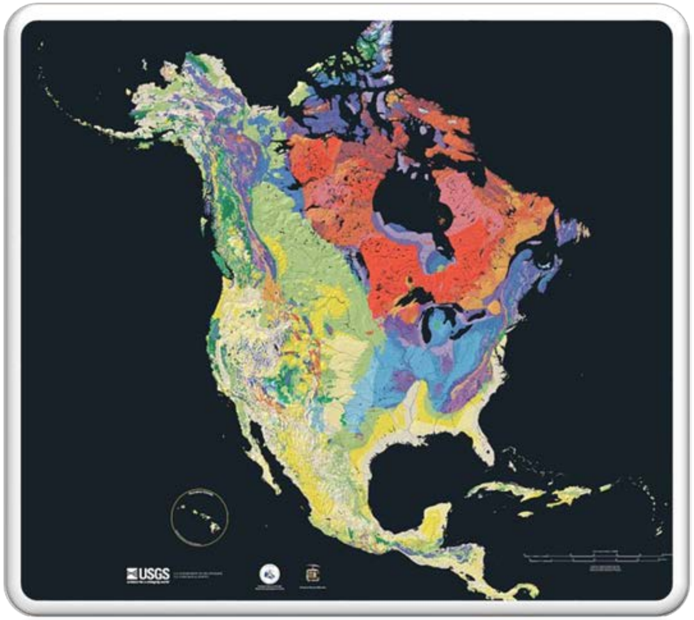
Long term

- Create protocols for map editing and cartography
- Create processes to expedite field data entry, map production, publication, etc.
- Build/code tools to increase efficiency
- **Assign compilation geologist**; begin work on compilation
- Routine database and system maintenance

Data sharing helps everyone make better decisions

Geologic information is used by nearly every component of the government

- Agriculture
- Commerce
- Defense
- Education
- Energy
- Health and Human Services
- Homeland Security
- Housing and Urban Development
- the Interior
- Labor
- State
- Transportation
- the Treasury



Possible data-sharing methods

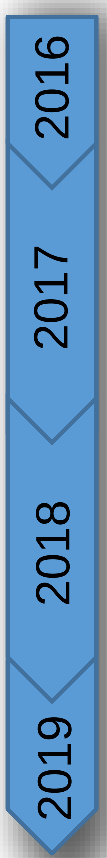
- Web feature service TBD
- GeoSciML, http://www.geosciml.org/
- GeoSciML-Portrayal for WFS and WMS
- *Other?*

Considerations for pilot

- Transfer speed
- Scalable – some data sets will be huge
- Allows use of production tools
- Ease of use
- Allows simplified geologic terms to be transferred
- Compatible with single-map NCGMP09 and other database models
- *Others?*

A way forward

Many heads are better than one



- 2016 DMT: Initial input and discussion
- 2016-17 NCGMP09 workgroup meetings
  - **Database model** development and testing with 2 data sets
- 2017 DMT: Update on database model
- 2017-18 NCGMP09 workgroup meetings
  - **Data-sharing protocol** development and testing
- 2018 DMT: Update on pilot data-sharing protocol
- 2018-19 Documentation
- 2019 DMT: Update on project
- 2019 Code and models will be posted to EPA's repositories