Migration to a Modern Relational Database: A Repository’s Rocky Road to Improved Geologic Inventory Management

Speaker: Kurt Johnson
State of Alaska
Department of Natural Resources
Division of Geological & Geophysical Surveys
Geologic Materials Center
Overview

Tale of Two Repositories

Impacts of Change
Pre-move Factors
Database Value to GMC Systems
Database Nuts and Bolts
Reflections & Source Code

Alaska Spring 2015
Origin
(Eagle River Fish Hatchery)

Challenges
- Sprawling site
- Warehouse over capacity
- 65 storage areas
- Winter
Destination
(former Anchorage Sam's Club)

Site advantages
- Easier client travel
- Climate controlled
- Renovated infrastructure

Staff: Kurt Johnson, Jean Riordan  IT: Ken Woods
Interns: Alexandra Busk, Walter Zimmerman
Volunteers: Robert Ravn, Robert Blodgett, Joan Taylor
Facility Changeover

Problems improved

- Work environment
- Sample stability
- Viewing space
- Logistics
Impacts of Change

Transformation

- It’s stressful and a lot of work
- Integrating new concepts is hard
- Added complexity needs IT support
- Likely loss of in-house control

Walmart only has 142,000 items to track and rocks are much heavier
Dealing with Change

Recommendations

• Conceptual appropriation is good
• Give a lot of lead time
• Initial data organization may be easier in Microsoft Excel and Access
• Facility staff is responsible for data
• Might be easier to fix data in new database
Information Considerations

Live tracking between two facilities
Inexpensive
Staff can manage data integrity
Spatial architecture (GIS)
Desktop to enterprise flexibility
Database Selection

**PostgreSQL** – the world's most advanced open source database
- Zero licensing cost
- Easy to maintain

**PostGIS** – open source spatial extension
- Extensive support for GIS queries and tools
- Allows map-based location SQL queries
Our PostgreSQL Setup

Modest hardware

Full backups every 2 hours

Segregated database environment

- gmc_prod – production database
- gmc_dev – development database
  - Test without fear of breaking things
  - Cloned from production every night
Scope of Collections

3063 Energy Wells
- 16 million representative feet

272 Mineral Prospects
- 2,158 boreholes
- 532,000 representative feet

450,000 Surface Samples

230,000 Processed Slides

140,800 Boxes on Shelves

607,600 Inventoried Items
Move Preparations

Find collections
- Inventory all sample boxes
- “Box with barcode on shelf”

MS Access transition
- Centralize data
- Standardize fields & coding

Data migration plan
- Develop schema
- Load PostgreSQL

Audit entire facility
The Move

In total it took:
31 working days to shift
102,741 sample boxes weighing
2.5 million pounds stacked on
2,522 pallets moved in
142 trucks with
0 injuries
Database Value to GMC Operations

Warehouse hand scanners
Inventory web browser interface
- External facing – public searches
- Internal facing – GMC staff

In development
- Client tracking
- Sampling history
- Scanner upgrade
Warehouse Hand Scanners

Benefits

- Interacts directly with database via Wi-Fi
- Retrieves sample box metadata
- Small updates on inventory
- Carries out audits
- Water resistant and can survive drop onto concrete floor
- Great for playing Star Trek
Custom Web Inventory Interface

All search results stored in URL

http://maps.dggs.alaska.gov/gmc/
Spatial Searches

**Web interface example**

- User delimits search area
- Point or polygon results
- Integrates spatial and text searches

User bounding box search near ANWR, Alaska
Internal Browser Interface

Features for staff
- Login
- QA list
- Import data
- Edit data
- Box move log
- Box audit
Database Nuts and Bolts

Objectives

- Less complex schema
- Flexible data storage
- Easier SQL
- Faster transactions
GMC Schema
(“Ever ting irie mon!”)
Recursive Data and Triggers

Advantages
- Flexible hierarchical data structure
- Automates changes to data structure

Key concepts
- Self-referential table
- Branched records
- Root position (seed)

GMC example
- Move cabinet
- Trigger cascades changes
- Human readable path is updated

GMC “containers” (not shelves)

Containers hold samples

- Site
- Section
- Stack
- Column
- Shelf
- Pallet
- Box
Data Loading

Temporary database table
- Flexible import format for arbitrary data
- Perform initial cleanup
- Automatically dropped

JSON data type
- Stores arbitrary data
- Helps parse data to final tables

User defined functions
- Can be written in numerous languages
- Easier SQL
Array and Enum Data Types

Advantages
- Fewer tables in schema
- Easier SQL
- Faster searches
- Consistent cataloging

GMC “keyword”
- Enum field type contains the “master list”

GMC “keywords”
- Array field stores selected options

GMC keywords example
processed, core, palynology, slide

THE ULTIMATE GROCERY LIST

PRODUCE
- 1 avocado
- 1 lemon
- 1 tomato
- mint
- 2 bananas
- 1/2 cup spinach
- 1/2 cup kale

PROTEIN
- 6 eggs

CANNED
- 1 can black beans
- 1 can coconut cream
- 1 can salsa

DAIRY
- 1 bag shredded low-fat cheddar cheese
- 1 quart almond milk

FROZEN
- 2 100-gram packs frozen açai pulp or purée

SWEETS
- 1 dark chocolate bar
- 1 package shredded coconut

PANTRY
- salt
- paprika
- honey
- cocoa powder
- hemp seeds
- coconut oil

GRAINS
- 1/2 loaf multigrain bread
- 1 package whole-wheat tortillas
- 1 package granola
- 1/2 cup quinoa

Shopping options (master list)
Range Data Types

Advantages
• 16 SQL lines become 1 line
• 2 database fields become 1 field
• Faster searches

GMC “intervals”
• Top and bottom depths define one segment
• Interval end point depths are inclusive

Down well sample intervals intersecting a formation pick
Advanced Pattern Matching

A regular expression (regex) is a sequence of characters that define a search pattern

Advantages

• Finds variations on names
• Cleans up name variations

Outcrop sample number example

“99 Mu-080a” + regex = "{99,Mu,080a}"

Regex example

'^(\d{2,4})-0,1([A-Z]{1,3})-0,1([\d\-]*A-Z.]*)$'
Database Transactions

Bundles multiple query steps into a single all-or-nothing operation that can be cancelled by user.

Advantages

- All pending changes can be rolled backed out or committed
- Very useful for debugging SQL scripts
Database Thoughts

Advice

• Modest normalization
• Generic table and field names
• Prepare to modify schema

Repository Truisms

• If it can go wrong, it already has gone wrong
• You can’t data model a geologist’s mind
• Don’t let perfection be the barrier to progress

Extreme 10-ton data loading tool
What we can share

Source code for
- Database
- Web interface
- Hand scanners

Challenges
- Not a boxed software product
- Lots of setup steps
- Stuffed into a State of Alaska template
- Schema boutique terms like "dggs_id"
Questions