Data Management, Data Delivery, and Standards

STEPHEN M. RICHARD

Data Rescue And Preservation Workshop
Utah Geological Survey, Salt Lake City Sept. 27, 2017
Outline

- Background
- Data management systems
- Data delivery
- Some example systems
- Standards
- Some conclusions
Data Preservation costs money

- Data are generally acquired with the purpose of solving some problem
- When the problem is perceived as solved, preserving the data is an afterthought.

OUR COMPETITION:

- Medical care
- Economic assistance
- Prisons
- Defense spending

- Highways, Dams, Airports
- Sports facilities
Start with WHY

- Society is better off if decisions are made based on sound scientific evidence
- The standard of living will improve if we can be more efficient and sustainable in use of natural resources
- It’s wasteful to acquire the same data over and over again
- It’s really cool to know that giant lizard-like creatures walked around my neighborhood 100 Million years ago
- Government open data rules
- It’s my job
What do users need?

- Hydrology, Engineering, Hazards, Environment:
  - Physical and chemical properties of the rocks
  - Gridded data to feed into computer models

- Mineral Exploration
  - Recognize map patterns—associations of rock type, age, structures, and documented mineralization/alteration

- Practical applications require detailed, site-specific local information
  - Typically at scales that geological surveys cannot provide
HOW?

- Paper records
- Institutional memory (People)
- Files on computers
- Databases
Data Access: files

Files – Tiered categories for data structuring
- Unstructured
- Structured
- Standardized (format, content)

Delivery – online repositories
- Package data items and metadata
- Management of these is agnostic about content (e.g. iRODS)
- DataOne data packages https://releases.dataone.org/online/api-documentation-v2.0/design/DataPackage.html
Databases

- Provide granular access to observations and feature descriptions
- Requirements vary widely and change over time within an organization
- Hard typing / Soft typing closed world/Open world
- Complexity is expensive to maintain
Data management database

- Updating, inserting QA/QC
- Reports, complex query capabilities
- Access controls
- Normalized—each fact in one place
- Controlled vocabularies
- Foreign key, uniqueness constraints
Data delivery

- Simplicity for user: no lookups, human-readable
- Performance – deliver quick
- Clear purpose

Archive

- Longevity-- can it be read in 10..20...50 years?
- Simple format: ASCII text, simple syntax
- Metadata bundled with content
- Error checks for detecting corruption
NoSQL, ‘document’ data bases

- Store ‘Documents’, typically structured, e.g. JSON
- No fixed schema
- Open world, very flexible
- Requires construction of indexes designed for purpose

Strabo—research project developing system for field data acquisition using Neo4j.
https://strabospot.org/
https://neo4j.com/
What Makes Data Usable

- I can’t use it if--
  - I can’t find it
  - I can’t get it
  - I don’t trust it

- Documentation is essential
  - Cost goes up as target community expands
Documentation

- Good documentation requires domain expertise AND catalog expertise.
  - Know details of how a given data set was created
  - Recognize facts that won’t be obvious outside of creator community
  - Understand standards

- Has to be created as close to the source as possible
  - Build into data acquisition workflow
  - ‘Backfilling’ is very expensive
Service Architecture

Data

Interface

Web Communication via interchange formats

Interface

User applications

User applications

User applications

User applications
Interchange format

- A way to structure the content of a file
  - Enables users can understand where information is located in the file

- XML, JSON, CSV, NetCDF: low level (syntactic)
- GeoSciML, WaterML, CF: includes semantics—labels with meanings (fields)

These are the actual basis for interoperability
Linked data

- Basically think of the web as a database
- Everything has a URI that can be resolved
  - IGSN—identify physical samples
  - DOI—identify publications and datasets
  - HTTP URI—identify anything
- Machine actionability—
  - Communicate to automated processes the formats available
  - Identifier schemes need to define scope and canonical representations
Examples: GeoSciML

- GeoSciML: comprehensive, complicated
  - GeoSciML lite—simplified, like NCGMP09
  - http://geosciml.org/ OGC standard
  - Provides schema for detailed rock and physical property description

- Uses:
  - OneGeology
    http://portal.onegeology.org/OnegeologyGlobal/
  - European INSPIRE services
Examples: NGDS

- National Geothermal Data System/USGIN
  - Simple flat file formats based on data actually in people’s databases
    [http://schemas.usgin.org/home/](http://schemas.usgin.org/home/)
  - Metadata conventions
    [http://usgin.github.io/usginspecs/MetadataRecommendationsGeoscienceResources_v1.2.htm](http://usgin.github.io/usginspecs/MetadataRecommendationsGeoscienceResources_v1.2.htm)
  - Services implemented by many surveys
  - Github for developing new content models
    [https://github.com/usgin-models](https://github.com/usgin-models)
  - Easily adapted for JSON or CSV interchange formats
Examples: CSV interchange formats

- EarthCube GeoWS
  - CSV interchange formats for various kinds of geoscience data
    https://www.earthcube.org/workspace/geows-geoscience-web-services/geows-web-services-summary

- CSV on the web, Data on the Web
  - W3C working group, developing specifications for text formats for tabular data
    https://www.w3.org/2013/csvw/wiki/Main_Page
    https://www.w3.org/TR/2017/REC-dwbp-20170131/
Data Type Registry

- Extend MIME-type concept to specify schema and semantics
- Defines vocabulary of entities and properties
- Binding of properties on entities
- Encoding scheme
### Data Type (Content model)

**Match fields to registered property**

<table>
<thead>
<tr>
<th>specie_name</th>
<th>location_name</th>
<th>latitude</th>
<th>longitude</th>
<th>resource_id</th>
<th>trait_id</th>
<th>trait_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micromussa amakus</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>40</td>
<td>40</td>
<td>Ocean basin</td>
</tr>
<tr>
<td>Micromussa amakus</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>40</td>
<td>41</td>
<td>Zooxanthellate</td>
</tr>
<tr>
<td>Micromussa amakus</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>10</td>
<td>96</td>
<td>Wave exposure preference</td>
</tr>
<tr>
<td>Micromussa amakus</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>10</td>
<td>97</td>
<td>Water clarity preference</td>
</tr>
<tr>
<td>Homophyllia bower</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>40</td>
<td>40</td>
<td>Ocean basin</td>
</tr>
<tr>
<td>Homophyllia bower</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>40</td>
<td>41</td>
<td>Zooxanthellate</td>
</tr>
<tr>
<td>Homophyllia bower</td>
<td>Great Barrier Reef (GB)</td>
<td>-17.832374</td>
<td>146.821289</td>
<td>38</td>
<td>65</td>
<td>Abundance GBR</td>
</tr>
<tr>
<td>Homophyllia bower</td>
<td>Global estimate</td>
<td></td>
<td></td>
<td>24</td>
<td>77</td>
<td>IUCN Red List cat</td>
</tr>
</tbody>
</table>

**‘Data type’: collection of properties about some entity of interest**
Use Scenario

Data sets in the catalog are mapped to registered data type.

I need data for samples that includes rock type and property.

entity

Data type registry

catalog
Standards

- Product that ships in volume
- Adoption
  - Requirement
  - Makes life easier– applications that use standard
  - It makes or saves money
- Counter factors
  - Science user market is small (no volume)
  - Research inherently seeks to try new things
Possible approach

- Start small: specific user group, specific problem, clear scope
- Document how it works
- Test under multiple setting and demonstrate utility in production
- Marketing is essential
- Provide support—documentation, working examples, training
Key considerations

- **Goal:** return on investment for data preservation and rescue
  - Benefit society

- **Challenges**
  - Small market, nugget effect
  - Data are public funded– paid forward
  - Technology is moving target
Conclusions

- What to preserve
  - What people are using
  - What can’t be replaced

- Sustainability
  - Have to demonstrate value/utility/impact
  - Seek economy of scale—consortium of surveys?
  - Cost has to be recognized and addressed
  - Professional recognition for data stewardship
Thanks
continuum

- Products -------------------------
  - Relatively static
  - Strong imprint of authors
  - Relatively small volume

----------------------

- data stream
  - Dynamic, constantly changing
  - Instrument source, largely automated, impersonal
  - Can be large volume