Perspectives on Data

A look at practices and procedures popular now

- Discussing Open Core Data
- Practices and Procedures of value
- Speculation

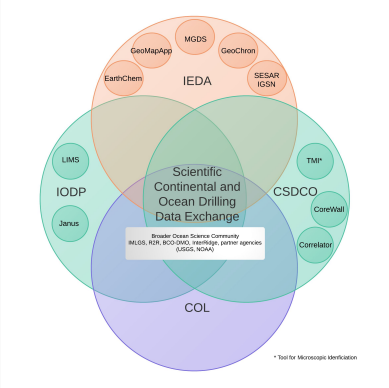
Open Core Data

http://opencoredata.org

Open Core Data is an infrastructure focused on making data from scientific continental and ocean drilling projects semantically discoverable, persistent, citable, and approachable to maximize their utility to present and future geoscience researchers.

The specific benefits that Open Core Data will bring to the community include:

- Integrating data management systems and services from multiple facilities, adding scientific value and economies of scale;
- Improving scientific drilling data discoverability and reuse through integration with evolving data infrastructures, augmenting existing domain-specific data systems (e.g. Neotoma, MagIC, EarthChem, PBDB, dbSEABED, GPlates) with scientific drilling data;
- Capturing and integrating PI-generated, post-moratorium scientific drilling data;
- Providing standards-based interoperability for tools to visualize and analyze scientific drilling data;
- Promoting and facilitating a Geoscience community of practice in data publication and citation;
- Providing a scalable resource that other communities and facilities could employ in the future (e.g. ANDRILL, ICDP, MGG-funded marine core repositories).



CSDCO Continental Scientific Drilling Coordination Office

IEDA INTERDISCIPLINARY EARTH DATA ALLIANCE

JOIDES Resolution Science Operator

CONSORTIUM FOR Ocean Leadership DISCOVERY • UNDERSTANDING • ACTION

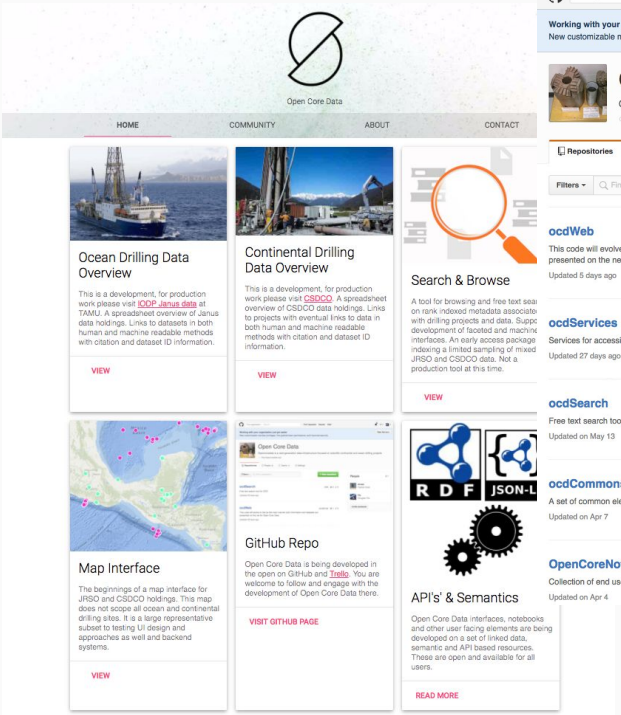
Open Core Data: Status

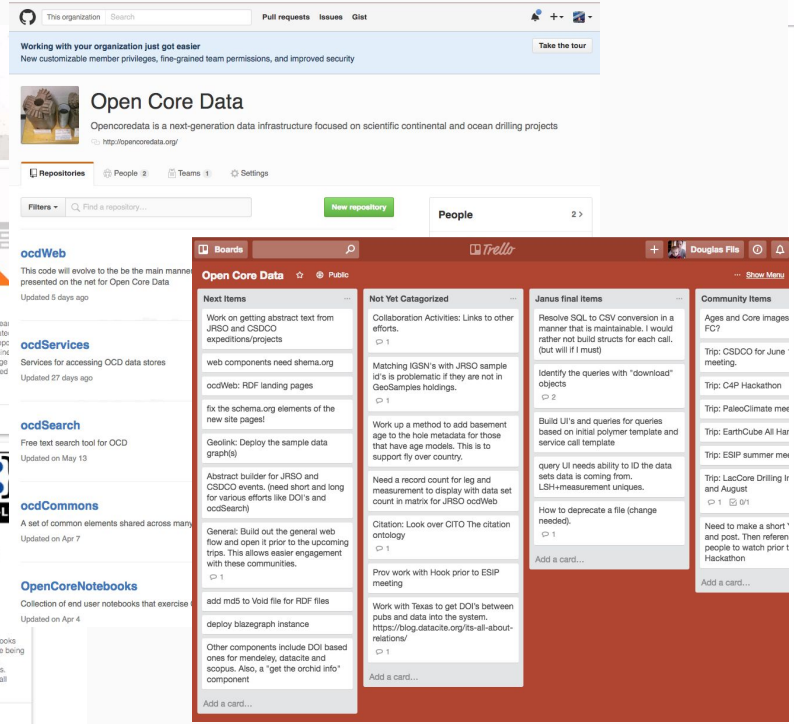
- Phase 1 development started in 2015 with a supplement to the Interdisciplinary Earth Data Alliance (IEDA)
- Current state shown at opencoredata.org. All code is open source, available at: github.com/OpenCoreData.
- NSF has informed the PIs that the Open Core Data Geoinformatics proposal will be funded in full.
- Initial work focuses on 4 major themes:

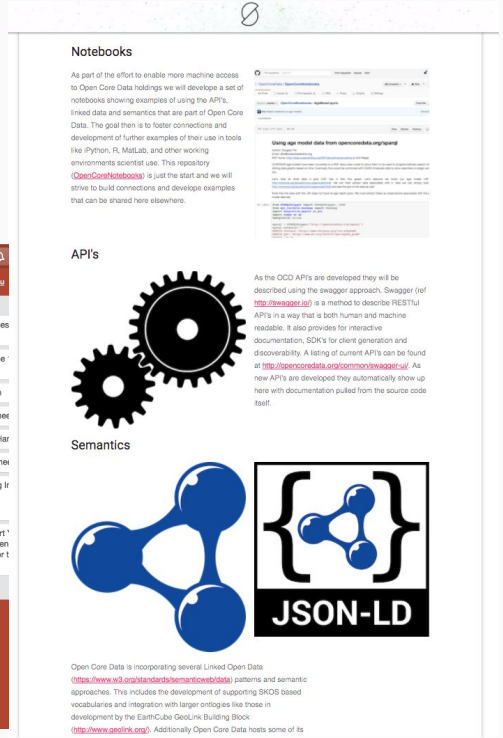
Migration	***Patterns and Models***	***Access***	***Discovery***
Initial work on moving data from JRSO and CSDCO holdings ~ 20K datasets with associated metadata so far	Exposing data and metadata in standards-based methods Using multiple formats to maximize human and machine access to data sets Examples: Schema.org RDF (GeoLink and others) CSV for the Web JSON-LD	Focus on both human and machine access. Integrating citable data (via DOIs) into science tools like iPython and others.	*Enhanced semantics* utilizing output from GeoLink (EarthCube Building Block) and other vocabularies *Linked Open Data* structures for machine indexing *Provenance and Citation* enhancement utilizing EarthCube and ESIP Federation outputs

Open Core Data: tour

A quick tour of Open Core Data with a look at the Linked Open Data, API and notebook plans.
http://opencoredata.org/ https://trello.com/b/dHxNEnCN/open-core-data https://github.com/OpenCoreData







Practices and Procedures

A survey of tools and approaches that facilitate data use and discovery.

Said another way…. Things you could do to make my life as a geoinformatics person easier.

- Computing Tools
- Unique ID’s
- File Formats
- Tools and Resources

Computing Tools

- R for Science https://ropensci.org/ http://r4ds.had.co.nz/ https://www.r-project.org/
  - Neotoma R: https://github.com/ropensci/neotoma
  - Neotoma API's http://api.neotomadb.org/doc/use

- iPython (Jupyter) https://ipython.org/ http://jupyter.org/
  - Good support at github: https://github.com/OpenCoreData/OpenCoreNotebooks/blob/master/OCDServiceNotebook_1.ipynb
  - Gplates examples: http://portal.gplates.org/ipython/
  - Anaconda https://www.continuum.io/why-anaconda

- OpenRefine http://openrefine.org/

- A wide range of tools (free and commercial) Excel, MatLab, etc. Even things like MongoDB, MySQL or personal data storages systems could be part of your local data architecture.

This is a huge and very personal aspect. Your workflow and tools are built out of your experiences and environment and are rarely going to result in the exact same solutions for everyone.

API's is a big topic.. I'm kinda skirting that in this version of the talk

Unique ID's

I am personally seeing quite a convergence in the Unique ID space. These include:

IGSN's for samples http://www.igsn.org/ and implementors like SESAR (http://www.geosamples.org/)

Orcid's for People (go get yours now while I am giving this talk, visit: http://orcid.org/)

Documents and Publications (DOI's)

Groups like DataCite (https://www.datacite.org/) support DOI's for data while more well CrossRef, etc for DOI's on publications.

Many services offer DOI capacity like Figshare, GitHub, IEDA, etc.

Figshare vs domain specific hosts

Is there a conflict between DOI's (with dx services) and Linked Open Data URI's ???

Research Institutes (http://www.re3data.org/)

COPDESS (http://www.copdess.org)

File Formats

What could be more boring that file formats? Maybe we should talk about the virtues of journaled file systems!

Still… some simple choices can help long term.

- JSON-LD (JSON is popular, JSON LD (linked data) implements a method to include a “context”)
  - Context? Simply a set of links, terms, definitions that define data types and provide descriptions
  - http://json-ld.org/
- CSV for the Web
  - https://www.w3.org/2013/csvw/wiki/Main_Page

More complex solutions exist like HDF5 and the upcoming Feather. Also there are more domain vertical solutions like LiPD for paleoclimate.

Tools and Resources

Open Science Framework https://osf.io/

Protocols https://www.protocols.io/

Mendeley https://www.mendeley.com

Github (or just git) https://github.com/

Figshare https://figshare.com/

Community Resources (IEDA, DataOne, Dryad, Neotoma, PaleoBioDB, ICDP, etc)

Center for Open Science https://cos.io/

Mozilla Science Lab https://science.mozilla.org/

Commodity resources: Google Drive, DropBox, Evernote, Citrix, Zoho… on and on..

Speculation

A wonderful title that allows me to guess and say things without worrying about finding this document 5 years from now.

- Greeting card thoughts
- Aspirations of data
- Structured and Unstructured data

Always see the exit! (remember Wild Bill Hickok)

Regardless of what tools, formats, or online environments you select always look for how to get out with all your stuff early and often. Select tools and environments that support easy export. Data Liberation!

Track Provenance (get it?)

The use of W3C Prov approaches for tracking samples, events, data, etc is coming. (I can feel it….. Which is why this section is called “speculation”)

Aspirations of data and [Un]Structured Data (2 for 1 slide)

- Integration
  - But how? Not sure, but the more you can describe the column, the units and the define “what this column is” the better. The machine AI’s are coming.. “So say we all”

- Semantic description
  - Fine for structured data.. (which is good). Look for winning vocabularies
    - Schema.org (http://schema.org)
    - GeoLink (http://geolink.org) Ok.. maybe not a “winner” but I am on this project.. So it’s here!
    - Other voc work like ODM2 https://github.com/ODM2/ODM2/wiki
  - Linked Data (what really will LOD do?)

- Machine inspectable data… give your data enough hints that the machines can go wild.. (my new motto)

My “guess” is that combining semantic data with unstructured data in a domain specific scope is the new black. It allows large collections of unstructured data to both benefit and influence smaller more focused structured data holdings.

Thanks

Douglas Fils
@fils
dfils@oceanleadership.org
http://orcid.org/0000-0002-2257-9127

Part of my mission and my interest is to see how to make data work for scientist.

Engage me, contact me, let me know your interest in scientific drilling data and I will do what I can to make things work for you and others.