Implementation of Open-World, Integrative, Transparent, Collaborative Research Data Platforms: the University of Things (UoT)

Prof. Peter Fox
(pfox@cs.rpi.edu, @taswegian, #twcrpi, ORCID: 0000-0002-1009-7163)
Tetherless World Constellation Chair, Earth and Environmental Science/ Computer Science/ Cognitive Science/ IT and Web Science
Rensselaer Polytechnic Institute, Troy, NY USA
And the Deep Carbon Observatory Data Science Team
CGA, Harvard, April 27, 2018
What to expect...

• Inevitable context, history + perspective
• Deep Carbon Observatory (Integration and Collaboration)
  – Data Science Platform for an international science community
  – Lots of RED, <WHITE> and BLACK
• data.rpi.edu V2 (Integration and Transparency)
• Where we are headed
  – Integration, Transparency and Collaboration
  – University Infrastructure for Data Science
Scientists – actually ANYONE - should be able to access a global, distributed knowledge base of scientific data and information that:

- appears to be *integrated*
- appears to be *locally* available
- is in a language (written, programming, or science) that is *understandable* and can be shared

**Data intensive – volume, complexity, mode, scale, heterogeneity, … in an OPEN WORLD**
Deep Carbon Observatory (DCO) …

• “We are dedicated to achieving transformational understanding of carbon’s chemical and biological roles in Earth.”

```
<table>
<thead>
<tr>
<th>TIME</th>
<th>2009-2012</th>
<th>Mid 2012-2013</th>
<th>Late 2012-2018</th>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCO Program Secretariat Established and Research Begins</td>
<td>Internal Engagement and Data Science Infrastructure Development and Implementation</td>
<td>External Engagement and Data Science Initiatives Launched and Research Project Activity Continues</td>
<td>Reporting and Synthesis Year</td>
<td>Dissemination Year</td>
<td></td>
</tr>
</tbody>
</table>
```

Building toward Synthesis and Dissemination

www.deepcarbon.net
• “Enable DCO team leaders to create new groups and associate a number of content types --- documents, discussions, blog posts, tasks, links, and bibliographic entries --- with the group, as well as simple event management (a private event calendar for the group) and embedding of external services (e.g. and esp. Google Calendar)” … more… (data, publications, projects)… a Knowledge Network … and a Virtual Organization (> 1000 people)
Producers | Consumers

- Data
  - Creation
  - Gathering

- Information
  - Presentation
  - Organization

- Knowledge
  - Integration
  - Conversation

Context

✓ Ecosystem metaphor
DCO Data Science Platform

DCO Object Registration and Deposit

Share

Deep Carbon Observatory
Data Science and Data Management Portal

Join

DCO Research Community Network

DCO Object Deposit

Knowledge Network

DCO Research Network

Register Metadata

Metadata

Dataset

Upload Raw Data

Title

Author

Author Email

Licence

Subject

Keyword

Data Type

GHS - Handle.net

CKAN

VIVO
Science Network of Things (Objects)

DCO Data Science Platform: Interaction of Architectural Components

DCO-Portal
- Drupal Application
- MySQL
- DCO Metadata Porting
- DCO Metadata Management
- Object Registration
- Object Discovery & Retrieval

info.deepcarbon.net

DCO-VIVO
- VIVO Application
- RDF
- DCO Dataset Deposit

DCO-ID (Handle System)
dx.deepcarbon.net

DCO-CKAN
data.deepcarbon.net

2012
DCO-ID as a Global Identifier
All information is linked and traceable!

Apatity and Space Research Institute

DCO ID
http://dx.deepcarbon.net/11121/5919-5240-5311-2220-CC

Deep Energy Community

DCO ID
http://dx.deepcarbon.net/11121/1677-5029-8638-9165-CC
Dataset Title
Noble gas isotope abundances in terrestrial fluids

DCO ID
http://dx.deepcarbon.net/11121/4317-8058-4791-8747-CC

Description
Global Data Base on isotopic composition of helium and other data for terrestrial fluids was compiled earlier by the Russian research group (B. Polyak, E. Prasolov, I. Tolstikhin, L. Yakovlev) and then was modified within the frame of this project supported by the Sloan Foundation. The DB is compiled as the Microsoft Excel spreadsheet. There are a few tables for regional records and a table for compiled total records of helium. In the dataset there is also an atlas document and a description document. Each MS-Excel file contains two spreadsheets: (1) Master sheet – contains all the compiled data (DATA) (2) References sheet – is the list of the data sources (REFS)


Data set accessed on (date) at http://dx.deepcarbon.net/11121/4317-8058-4791-8747-CC
• Knowledge network – implements both the collaboration and the integration, reporting implements the transparency – It’s being USED...
• Many means of population
  – User generation
  – Machine generation
• Contributing these enhancements back to open-source communities (CKAN, VIVO)
There’s more – Jupyter notebooks on top

And this: https://news.rpi.edu/content/2018/04/23/applying-network-analysis-natural-history
<table>
<thead>
<tr>
<th>Dataset Catalogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division of the Chief Information Officer Collections (1)</td>
</tr>
<tr>
<td>School of Architecture Collections (5)</td>
</tr>
<tr>
<td>Special Collections (1)</td>
</tr>
</tbody>
</table>

**Keywords**

- 2009 (1306)
- 5 (1)
- AAT (1)
- ACRES (118)
- AFS (118)
- ANSS (2)
- AQ (1)
- AQS (118)
- ARRA (1)
- Ability to speak English (1)
- Communications & Middleware Technologies, DotCIO, RPI (1)
- Experimental Media and Performing Arts Center (EMPAC), RPI (1)
- Horsford, Eben (1)
- Newberg, Heidi Jo (1)
- Yanny, Brian (1)
- data.gov (1724)

**Toxics Release Inventory Data**

- Description: Toxics Release Inventory Data.
- Handle: 10833/335
- Date Issued: 2011-04-03
- Authors: data.gov
- Contributors: data.gov
- Catalogs: Linking Open Government Data Collection
- Keywords: industries; Toxic Release; pollution prevention; Chemical Release; TRI State Data; section 313; 2009; Toxics Release Inventory; TRI; Chemical; community right to know; Toxics Release; TRI 2009 Data; NAICS; EPCRA; Hazardous; industry; Toxics; EPCRA 313; Toxic Release Inventory; source reduction; Toxic; Right to Know; Form A; TRI Reporting; facilities; toxic chemical release inventory; PBT; Chemical Pollution; TRI Data; toxic chemicals release inventory; community; Form R; waste management; facility; Chemicals; TRI Preliminary Data; TRI Report

**EPA FRS Facilities Combined File CSV Download for the State of Utah**

- Description: The Facility Registry System (FRS) identifies facilities, sites, or places subject to environmental regulation or of environmental interest to EPA programs or delegated states.
- Handle: 10833/1478
- Date Issued: 2011-03-16
- Authors: data.gov
- Contributors: data.gov
- Catalogs: Linking Open Government Data Collection
- Keywords: NCDB; FRS Data; TRIS; environment; latitude; environmental interest type; minor discharger; SDWIS; RCRAINFO; SQG; federal datasets; STATE; public health; major discharger; Transporter; TRI; facility site; affiliation type; DUNS; FRS; parent organization; program system; PCS; federal data download; national priorities list; longitude; LQG; NPDES; NEI; NAICS; organization; toxic release; ACRES; SSTS; regulated sites; RCRA; NCES; sites; SIC; facilities; Superfund; CAMDBS; CERCLIS; facility registry identifier; EPA; ICIS; Brownfields; Air Minor; AFS; program system identifier; RMP; AQS; Air Major; facility; NPL; RFS; program system acronym; TSD

**1992 Toxics Release Inventory data for the state of New York**

- Description: Toxics Release Inventory Data.
- Handle: 10833/1268
- Date Issued: 2011-04-03
- Authors: data.gov
- Contributors: data.gov
- Catalogs: Linking Open Government Data Collection
- Keywords: industries; Toxic Release; pollution prevention; Chemical Release; TRI State Data; section 313; 2009; Toxics Release Inventory; TRI; Chemical; community right to know; Toxics Release; TRI 2009 Data; NAICS; EPCRA; Hazardous; industry; Toxics; EPCRA 313; Toxic Release Inventory; source reduction; Toxic; Right to Know; Form A; TRI Reporting; facilities; toxic chemical release inventory; PBT; Chemical Pollution; TRI Data; toxic chemicals release inventory; community; Form R; waste management; facility; Chemicals; TRI Preliminary Data; TRI Report
Braasch, Jonas

Positions

Associate Professor, School of Architecture, RPI, Rensselaer Polytechnic Institute 2012 -

Publications in VIVO
5 total

Co-author Network

Map of Science

selected publications

dataset

Audio file documenting performance of automated music improvisation system CAIRA at the Music, Mind, and Invention Workshop

MATLAB code for model with head movement and ambisonics decoding

Performance of a sensory substitution system to present music via touch

Support video documentation for audio/visual psychophysical experiment

Video documenting performance of automated music improvisation system FILTER
Internal transparency and integration
Thus…

- Integrative – semantics
- Transparent – semantics
- Collaborative – semantics
- Application integration – Yep – semantics

So… where are we headed?
Research-grade but not “University-grade”

- Adoption of RDA outputs/recommendations
  - Data Type Registry
  - Permanent ID Types
  - Dynamic Data Citation*
  - Scholix*
- Improvements to VIVO
- Science network of things

- CIOs approach
  - “We only run the applications we know how to run”

- Library (not a research library)
  - Helped to start
  - Hurt in University adoption (hope^)

* underway
^ New Library Director
Progress toward a University of Things

- pfox@cs.rpi.edu and the DCO Data Science Team
- @taswegian twcrpi

- http://tw.rpi.edu
- http://tw.rpi.edu/web/project/DCO-DS
- http://deepcarbon.net
Garden shed
Framework v. systems v. platforms

Rough definitions

- Systems have very well-defined entry and exit points. A user tends to know when they are using one. Options for extensions are limited and usually require engineering.
- Frameworks have many entry and use points. A user often does not know when they are using one. Extension points are part of the design.
- Platforms ~ arise from frameworks.
VIVO Extension: Shibboleth Single-Sign-On
VIVO Extension: Dataset deposit in attached data repository

Begin

Need DCO-ID?

- YES
  - Collect CKAN metadata & generate URL
  - Revise CKAN metadata
- NO
  - Generate & register DCO-ID (unique suffix, blank URL)
  - Review DCO-ID & CKAN metadata
  - Update DCO-ID (map the DCO-ID to CKAN URL)

YES

NO

Revise

- YES
  - Collect CKAN metadata & generate URL
  - Revise CKAN metadata
- NO
  - Update DCO-ID record
  - Add URL (to data in external repository)
  - Deposit in CKAN & generate URL to data

YES

NO

End

DCO-ID & DCO-ID metadata deposited DCO data or URL to external data

Object without data URL

- Includes multi-level metadata collection
- Includes persistent identifier (DCO-ID generation)
- Includes interaction with dedicated repository OR accepts third-party deposit details
We identify ‘everything’ = DCO-ID

- Two part: all objects are issued Handle’s, and all published objects are also issued DOIs
  - DCO issues Handles, registration number is 11121
  - We obtain DOIs from DataCite
  - If it is a person, we support ORCID, ResearcherID, ScopusID, eRA Commons, etc.

- You may see (note EPIC style identifier syntax):
  http://hdl.handle.net/11121/5676-3964-8313-5126-CC and
  http://dx.deepcarbon.net/11121/5676-3964-8313-5126-CC

- E.g. Adding bibliography is easy, just enter the DOIs, or paste a bibtex record, and we do the rest, same for people (ORCID, ResearcherID, etc.) -> open world – linked to other sources
Imported Metadata for Publication DOI: 10.1109/TGRS.2013.2262179

Publication Type
Select one

Title
Provenance Representation for the National Climate Assessment in the Global Change Information System

Publication Year
2013

Authors
Tilmes, Curt
Fox, Peter
Ma, Xiaogang (create this person in VIVO)
Ma, Xiaogang (Marshall)
Miyajima, N
Martinelli, G
Caramanna, C
Matsui, N
Smith, Aaron
Waple, Anne
Zednik, Stephan
Zheng, Jin Guang

Venue
IEEE Transactions on Geoscience and Remote Sensing
VIVO Extension: Retrieval of DOI metadata for publications from CrossRef
* Expedites entry of e.g. journal articles by retrieving metadata based on DOI
* Preserves author rank
Core and Framework Semantics - Multi-tiered interoperability

Mediation!

Data-level Semantic mediation: lower-level vocabularies applied to each data source for a specific science domain of interest

- Federal Repository
- Commercial Database
- Researcher Private Database
- Other Data Sources

Metadata, schema, data

2012