Progress in Open-World, Integrative, Transparent, Collaborative Science Data Platforms

Prof. Peter Fox (pfox@cs.rpi.edu, @taswegian, #twcrpi)
Tetherless World Constellation Chair, Earth and Environmental Science/Computer Science/IT and Web Science)
Rensselaer Polytechnic Institute, Troy, NY USA
And collaborators (esp. Deep Carbon Observatory and Global Change Information System)
ISWC 2013, Sydney Australia, October 24, 2013
What to expect…

• Inevitable context, history + perspective
• Systems v. Frameworks -> Platforms
• Science Data -> Data Science
• Integration, Transparency and Collaboration
• Deep Carbon Observatory (Integration and Collaboration)
• Global Change Information System (Integration and Transparency)
• What is next in the ecosystem!
What to expect...

• Inevitable context, history + perspective
• Systems v. Frameworks -> Platforms
• Science Data -> Data Science
• Integration, Transparency and Collaboration
• Deep Carbon Observatory (Integration and Collaboration)
• Global Change Information System (Integration and Transparency)
• What is next in the ecosystem!
RPI Tetherless World Constellation

tw.rpi.edu

- Government Data
- Health care/Life Sciences
- Environmental Informatics

Lots of RDF
Web Infrastructure
Scaling and Distributed
Query and Reasoning
AI, Rule reasoning
Visualizing SW ‘data’
Social SW, Sem-Drupal
Policy Lang, Events, Processes
Ontologies/Tools, …

Future Web
- Web Science
- Policy
- Social

Xinformatics
- Data Science
- Semantic eScience
- Data Frameworks

Semantic Foundations
- Knowledge Provenance
- Inference, Trust

Hendler
Fox
McGuinness

Luciano, Erickson
+ ~ 40 = Post-docs, Staff, Grad, UGrad

• Future Web
  • Web Science
  • Policy
  • Social

• Xinformatics
  • Data Science
  • Semantic eScience
  • Data Frameworks

• Semantic Foundations
  • Knowledge Provenance
  • Inference, Trust

Government Data
Health care/Life Sciences
Environmental Informatics

Lots of RDF
Web Infrastructure
Scaling and Distributed
Query and Reasoning
AI, Rule reasoning
Visualizing SW ‘data’
Social SW, Sem-Drupal
Policy Lang, Events, Processes
Ontologies/Tools, …
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**
In 2000-2001 the need for capturing and preserving knowledge in science data became very clear but the barriers were high.

In 2004 we started a virtual observatory project based on semantic technologies.

Use case driven – in solar and solar-terrestrial physics with an emphasis on instrument-based measurements and real data pipelines; we needed implementations.

We knew we also needed integration and provenance (but that came later).

We aimed to push semantics into our systems to build new ‘prototypes’ but we ‘failed’ ;-)
In 2004

- 2004 – OWL was a W3 recommendation!!
- Protégé 2.x and the Protégé-Java-OWL API
- SWOOP was a viable editor
- Jena and the Jena API were in good shape
- Pellet worked
- SPARQL was still a twinkle in the RDF working group’s eye (there were competitors… OWL-QL)
- Semantics were still the realm of computer scientists
Developing ontologies

- Use cases and small team
- Identify classes and minimal properties (leverage controlled vocab.)
- Review, vet, publish
- Only code them (in RDF or OWL) when needed (CMAP, ...)
- Ontologies: small and modular
Semantics between 2004 and 2009

- Ontologies were needed for data integration and provenance and mediation for data mining
- Protégé 3.x and then 4.0 came out
- SWOOP development was interrupted
- Cmap added OWL predicate support*
- SPARQL became a recommendation
- Triple stores exploded in use and capability
- Linked Open Data started to take off
- Pellet 2.0 came out
Working with knowledge

Expressivity

Implementability

Maintainability/Extensibility
Working with knowledge
Semantics between 2009 and 2013 (for us)

• Semantic data frameworks (SeSF)
• Substantial knowledge provenance work
• Data quality, uncertainty and bias representations and applications
• Applications:
  – Sea Ice, Carbon Observatory, Integrated Ecosystem Assessments, Global Change Information System, ocean.data.gov, energy.data.gov ....
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.
Core and Framework Semantics - Multi-tiered interoperability

Discipline-specific model(s)  
Semantic interoperability  
Dataproduct Generator  
Information/Science Apps

Semantic query, hypothesis and inference  
Query, access and use of data

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
Inter-disciplinary Data Visualization Apps

Integration Frameworks & Methodologies

Eco & other system Assessment Apps

Inter-disciplinary Data Visualization Apps

Integration Frameworks & Methodologies

Eco & other system Assessment Apps

Discipline-specific model(s)

Dataproduct Generator

Information/Science Apps

Core and Framework Semantics - Multi-tiered interoperability

People

Agency Policy Makers

System Scientists

Politicians

Integrated Applications

Decision-level semantic mediation: high-level vocabularies that facilitate policy-level decision-making

Application-level semantic mediation: mid-level vocabularies that facilitate the interoperability of system models and data products

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

Software, Tools & Apps

Data Repositories

Federal Repository

Commercial Database

Researcher Private Database

Other Data Sources

Metadata, schema, data

Semantic query, hypothesis and inference

Query, access and use of data

Multi-tiered interoperability

Query, access and use of data

Semantic query, hypothesis and inference

Semantic query, hypothesis and inference
Core and Framework Semantics - Multi-tiered interoperability

Integrated Applications

Mediation!

Inter-disciplinary Data Visualization Apps
Semantic interoperability
Integration Frameworks & Methodologies
Eco & other system Assessment Apps

Application-level semantic mediation: mid-level vocabularies that facilitate the interoperability of system models and data products

Software, Tools & Apps

Mediation!

Discipline-specific model(s)
Semantic interoperability
Dataproduct Generator
Information/Science Apps

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

Data Repositories

Mediation!

Federal Repository
Commercial Database
Researcher Private Database
Other Data Sources

Metadata, schema, data

People

Mediation!

Agency Policy Makers
System Scientists
Politicians

Decision-level semantic mediation: high-level vocabularies that facilitate policy-level decision-making
Inter-disciplinary Data Visualization Apps
Integration Frameworks & Methodologies
Eco & other system Assessment Apps

Integrated Applications
Inter-disciplinary Data Visualization Apps
Semantics interoperability
Integration Frameworks & Methodologies
Eco & other system Assessment Apps

Software, Tools & Apps
Discipline-specific model(s)
Semantic interoperability
Dataproduct Generator
Information/Science Apps

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

Decision-level semantic mediation: high-level vocabularies that facilitate policy-level decision-making

Application-level semantic mediation: mid-level vocabularies that facilitate the interoperability of system models and data products

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

Core and Framework Semantics - Multi-tiered interoperability
Data Science is …

• Doing science with someone else’s data …
Data Science is …

- Doing science with someone else’s data …
  - across datasets
Data Science is …

- Doing science with someone else’s data …
  - across datasets
  - with models
Data Science is …

• Doing science with someone else’s data …
  – across datasets
  – with models
  – multi-dimensional, multi-scale, multi-mode
Data Science is …

• Doing science with someone else’s data …
  – across datasets
  – with models
  – multi-dimensional, multi-scale, multi-mode
  – complex data types
Data Science is …

• Doing science with someone else’s data …
  – across datasets
  – with models
  – multi-dimensional, multi-scale, multi-mode
  – complex data-types
  – needing new analytic and visual approaches
Data Science is …

• Doing science with someone else’s data …
  – across datasets
  – with models
  – multi-dimensional, multi-scale, multi-mode
  – complex data-types
  – needing new analytic and visual approaches
• Especially in multiple “dimensions” (functional)
  – E.g. Detection/ attribution methods/ algorithms
  – Visual exploration
• Today – is collaborative, performed in a network, and web-based …
Ecosystem metaphor

Producers

Data

Information

Knowledge

Consumers

Context

Creation
Gathering

Presentation
Organization

Integration
Conversation

Experience
What else is in the ecosystem?

• Many elements, and we still do not have sufficient *information models* of how they inter-relate.

• But these are what enable data scientists to explore/confirm/deny their ‘hunches’ (hint for later).

Collaboration  Accountability  Identity

Explanation  Justification  Verifiability  Proof  Trust

Integratability  Citability

‘Transparency’ -> Translucency
What else is in the ecosystem?

- Many elements, and we still do not have sufficient *information models* of how they inter-relate.
- But these are what enable data scientists to explore/confirm/deny their ‘hunches’ (hint for later).

Collaboration  Accountability  Identity

Explanation  Justification  Verifiability  Proof  Trust

Integratability  Citability  ‘Transparency’ -> Translucency

‘Provenance’
Modern informatics enables a new scale-free framework approach

- Use cases
- Stakeholders
- Distributed authority
- Access control
- Ontologies
- Maintaining Identity
We are dedicated to achieving transformational understanding of carbon’s chemical and biological roles in Earth.

**TIMELINE**

<table>
<thead>
<tr>
<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>
</tr>
</tbody>
</table>

**Building toward Synthesis and Dissemination**
Deep Carbon Observatory (DCO) …

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

<table>
<thead>
<tr>
<th>TIMELINE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2009-2012</strong></td>
</tr>
<tr>
<td>DCO Program Secretariat Established and Research Begins</td>
</tr>
</tbody>
</table>

Building toward Synthesis and Dissemination
Decadal goals = Discovery science

Global community of ‘Carbon’ scientists contributing to Deep Earth Computer (data legacy) comprising:
Decadal goals = Discovery science

Global community of ‘Carbon’ scientists contributing to Deep Earth Computer (data legacy) comprising:

• Global Earth Mineral Laboratory
• Global Census of Deep Fluids
• Global Volcano Gas Emissions
• Global Census of Deep Microbial Life
• Global State of High Pressure and Temperature Carbon and Related Materials
• Global Inventory of Diamonds with Inclusions
“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…
DCO Data Science Platform

DCO Object Registration and Deposit

Share Knowledge

Register Metadata

Upload Raw Data

Allocate a universal accessible DCO-ID

DCO Research Network

Join Network

DCO Science Network

Metadata

• Title
• Author
• Author Email
• Licence
• Subject
• Keyword
• Data Type

Dataset

CDF

DCO Object...

DCO-ID

Research

Community

Network
DCO Data Science Platform

DCO Object Registration and Deposit

- Title
- Author
- Author Email
- Licence
- Subject
- Keyword
- Data Type

Register Metadata

Dataset

Upload Raw Data

DCO-ID Request

Allocate a universal accessible DCO-ID

CKAN

VIVO

DCO Research Network

DCO Science Network
DCO Data Science Platform
DCO Data Science Platform

Register Metadata

Dataset

- Title
- Author
- Author Email
- Licence
- Subject
- Keyword
- Data Type

Upload Raw Data

DCO Object Registration and Deposit

Share Knowledge

DCO Object Deposit

Join Network

DCO Research Network

Allocate a universal accessible DCO-ID

DCO Science Network

DCO Research Community Network
Comprehensive Knowledge Archive Network (used for data.gov, others...)

CKAN, the world's leading open-source data portal platform
CKAN is a powerful data management system that makes data accessible — by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

Feature Overview
- Complete catalog system with easy to use web interface and a powerful API
- Strong integration with third-party CMS's like Drupal and WordPress
- Data visualization and analytics
- Workflow support lets departments or groups manage their own data publishing
- Fine-grained access control
- Integrated data storage and full data API
- Federated structure: easily set up new instances with common search

Support and Hosted Solutions
CKAN is open source and can be downloaded and used for free. Users can also get hosting and support from a range of suppliers.
A full-time professional development team at the Open Knowledge Foundation maintains CKAN and can provide full support and hosting with SLAs.
VIVO - represents academic research communities
• Every person, organization, or other data entity in VIVO has a unique identifier
• VIVO enables the discovery of research and scholarship across disciplines at one institution or across many
• Records are both human-readable and machine-readable
• We’ve extended (yes, ontologies) VIVO to the science network – datasets, instruments, sites, etc.
• Feeding this back to VIVO
Welcome to the Handle System

The Handle System provides efficient, extensible, and secure resolution services for unique and persistent identifiers of digital objects, and is a component of CNRI's Digital Object Architecture. Digital Object Architecture provides a means of managing digital information in a network environment. A digital object has a machine and platform independent structure that allows it to be identified, accessed and protected, as appropriate. A digital object may incorporate not only informational elements, i.e., a digitized version of a paper, movie or sound recording, but also the unique identifier of the digital object and other metadata about the digital object. The metadata may include restrictions on access to digital objects, notices of ownership, and identifiers for licensing agreements, if appropriate.

The Handle System includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the resources. This information can be changed as needed to reflect the current state of the identified resource without changing its identifier, thus allowing the name of the item to persist over changes of location and other related state information. Some examples of applications that use HDL® identifier and resolution services as infrastructure are rights management applications, digital object registries and repositories, and institutional data preservation and archiving.
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

• You would see (note EPIC style identifier syntax):
  http://hdl.handle.net/11121/00-0123-4567-8901-1234-C and http://dx.deepcarbon.net/11121/00-0123-4567-8901-1234-C

• 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
DCO Data Science Platform: Interaction of Architectural Components

DCO-Portal
- Drupal Application
- MySQL
- DCO Metadata Porting

Object Discovery & Retrieval

Object Registration

DCO-VIVO
- VIVO Application
- RDF
- DCO Metadata Management

DCO-ID (Handle System)

DCO-CKAN
- DCO Data Repository

DCO Dataset Deposit
Schematic for Deep Carbon Observatory Data Flows

Data Management and Data Science Guidance - DCO-wide and Compatible with National and International Best Practices

Experiments -> Physics/Chemistry Models -> Sensor streams -> Sequencing -> Sample collection

- Standards based, Provenance captured
- Efficient generation, identifiers issued
- Query, access and use of existing data

Many means of generation
Schematic for Deep Carbon Observatory Data Flows

Data Management and Data Science Guidance - DCO-wide and Compatible with National and International Best Practices

Many means of generation

Experiments

Physics/Chemistry Models

Sensor streams

Sequencing

Sample collection

Global Census, Virtual Mineral Laboratory, ...

Standards based, Provenance captured

Efficient generation, identifiers issued

Query, access and use of existing data

DCO data infrastructure; identifiers, Catalogs

Existing Community Data Infrastructure; Identifiers, Catalogs

DCO-wide and Compatible with National and International Best Practices

GIDI

Others

EOS

EarthChem

MINDAT

Emission/Compositions

Metadata, schema, data ...

Global Census, Virtual Mineral Laboratory, ...

Data Repositories

Phy...
Schematic for Deep Carbon Virtual Observatory and Interoperability

Software, Tools & Apps

Data Repositories

Semantic mediation: physics, chemistry, mineral, emission data - ChemML,

Deep Energy/Life Applications

Physics/Chemistry Models

Res/Flux Applications

Semantic query, hypothesis and inference

Query, access and use of data

Semantic interoperability

GVP

MINDAT

EOS

EarthChem

Emission/Compositions

Metadata, schema, data... ... ...

Venting, apprehension, ...
Schematic for Deep Carbon Virtual Observatory and Interoperability

Integrated Applications

Software, Tools & Apps

Data Repositories

Discovery visualizations

Analytics and mining

Global Census, Virtual Mineral Laboratory, ...

Application-level mediation: vocabulary, mapping to science and data terms

Deep Energy/Life Applications

Physics/Chemistry Models

Res/Flux Applications

Semantic query, hypothesis and inference

Semantic mediation: physics, chemistry, mineral, emission data - ChemML,

GVP

MINDAT

EOS

EarthChem

Emission/Compositions

Query, access and use of data

Metadata, schema, data ...

Semantic interoperability

Semantic interoperability
Groups

<table>
<thead>
<tr>
<th>Title</th>
<th>Subgroups</th>
<th>Posts</th>
<th>Members</th>
<th>Description</th>
<th>Created</th>
<th>Leave</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCO Engagement Basic Group. Restricted</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2013-10-03</td>
<td></td>
</tr>
<tr>
<td>DCO Data Science Community Science Working Group</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>The DCO Data Science &amp; Data Management Community</td>
<td>2012-08-20</td>
<td></td>
</tr>
</tbody>
</table>

Your groups
- DCO Data Science Community
- DCO-DS and GVP Gas Emission Extension Boundary Activity Group

Most active groups
- DCO Data Science Community
  - 5 member(s)
  - 8 post(s)

Newest groups
- DCO Data Science Community
  - The DCO Data Science & Data Management Community
Using an integrated approach combining informatics, data science and life cycle data management best practices, the objectives of the DCO Data Science team are to establish a Deep Carbon Virtual Observatory (DCVO) for DCO embedded in community and agency data resource holdings; provide robust data infrastructure for DCO science, instrumentation, secretariat and engagement activities; enable scientific discovery via visualization and analysis in concert with the DCO data infrastructures; and advance educational aspects of data science and data management among all DCO participants.

Planned activities by the DCO Data Science include: a) Develop data methodologies to handle a global network of volcano monitoring stations; b) Create tools to process new streams of sensor data returned in near-real time for science and public engagement; c) Rapidly model new data that results from probing into the deep interior with a combination of drilling and experimental innovations; d) Encode and store the simulation results of computations in new temperature and pressure regimes of carbon physics/chemistry/biology; e) Capture data from new specialized laboratory instrumentation; f) Integrate data from different disciplines; genomic sequence data and increasingly complex metadata of life in extreme environments; and g) Innovate new data tools that are quickly and easily accessible to a wide audience, including tools to visualize and analyze large data holdings.

The primary expected product of the DC Data Science work will be a Deep Carbon Virtual Observatory and DCO data infrastructure that meets DCO data science and management research and education needs and requirements across the data life cycle. A DCO science network, facilitating routine collaboration. Secondary products of this work include providing a basis for an Earth Materials Data Infrastructure.

The DCVO allows discovery, dissemination and use of DCO data and is used by the DCO community and other researchers. Community Data Resource agreements and interface specifications are in place that leverage and enhance the significance of the community data resources. A DCO Science Network that routinely collaborates in research on--line. Visualization infrastructure and methodologies are in use for DCO. Data from DCO projects becomes as valuable as the research papers written about them. Data Science and Data Management education and training will have an impact on the way DCO researchers, particularly those in their early career, conduct their science and share data. The expected impact would be significant returns on resource investments made on generation or acquisition of DCO science data.

Research Activity Reporting

DCO activities are monitored regularly by the DCO Secretariat, Scientific Steering Committees, and the Sloan Foundation. In order for these activities to be documented, they must be registered and deposited in the DCO Data Portal. Using the links to the right, please enter information pertaining to this group’s activities and achievements relevant to DCO goals and research objects.

- Grant/Other Funding Proposal
- Honors and Awards
- Fieldwork
- Partnerships/Collaborations
- Instrumentation
- Outreach/Media
- Data-related activity

NOTE: The links above lead to services on the DCO Data Portal
A DCO science network: facilitating routine collaboration. Secondary products of
this network are used by the DCO community and other researchers. Community Data Resource agreements and
deposition of significant community data resources. A DCO Science Network that routinely collaborates in research
use for DCO. Data from DCO projects becoming as valuable as the research papers written about them. Data
are an impact on the way DCO researchers, particularly those in their early career, conduct their science and
on resource investments made on generation or acquisition of DCO science data.

Report an achievement or award:
- Grant/Other Funding Proposal
- Honors and Awards
- Fieldwork
- Partnerships/Collaborations
- Instrumentation
- Outreach/Media
- Data-related activity

Submit relevant content:
- News Feature
- Image/Graphic
- Video
- Audio
- Bibliographic Item
- Deep Carbon General Interest Item
- Dataset

NOTE: The links above lead to services on the DCO Data Portal.

NOTE: The links above lead to services on the DCO Data Portal.
A DCO science network: facilitating routine collaboration. Secondary products of this structure and is used by the DCO community and other researchers. Community Data Resource agreements and management of significant community data resources. A DCO Science Network that routinely collaborates in research use for DCO. Data from DCO projects becoming as valuable as the research papers written about them. Data on an impact on the way DCO researchers, particularly those in their early career, conduct their science and resource investments made on generation or acquisition of DCO science data.

**Report an achievement or award:**
- Grant/Other Funding Proposal
- Honors and Awards
- Fieldwork
- Partnerships/Collaborations
- Instrumentation
- Outreach/Media
- Data-related activity

**Submit relevant content:**
- News Feature
- Image/Graphic
- Video
- Audio
- Bibliographic Item
- Deep Carbon General Interest Item
- Dataset

**NOTE:** The links above lead to services on the DCO Data Portal.

---

**NOTE:** Deposited in the DCO Data Portal. Using the links to the right, please enter information pertaining to this group’s activities and achievements relevant to DCO goals and research objects.
International Meeting Report

On Sunday, 3 March 2013, more than 100 scientists in the Deep Carbon Observatory (DCO) network converged on the US National Academy of Sciences on Washington, DC’s National Mall. Having so many DCO scientists together...

Video: Carbon’s Untold Secrets

DCO scientists describe the significance of deep carbon and the many mysteries that the ten-year program hopes to resolve. Video courtesy of Chemical & Engineering News.
## Community Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Institution</th>
<th>DCO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>West, Patrick</td>
<td><a href="mailto:westp@rpi.edu">westp@rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/5900-7104-5112-6072-CC">http://dx.deepcarbon.net/11121/5900-7104-5112-6072-CC</a></td>
</tr>
<tr>
<td>Erickson, John</td>
<td><a href="mailto:erick14@rpi.edu">erick14@rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/2420-7841-6636-4617-CC">http://dx.deepcarbon.net/11121/2420-7841-6636-4617-CC</a></td>
</tr>
<tr>
<td>Wang, Han</td>
<td><a href="mailto:wang17@rpi.edu">wang17@rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/7999-9599-5031-4221-CC">http://dx.deepcarbon.net/11121/7999-9599-5031-4221-CC</a></td>
</tr>
<tr>
<td>Chen, Yu</td>
<td><a href="mailto:cheny10@rpi.edu">cheny10@rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/4310-7699-0054-1394-CC">http://dx.deepcarbon.net/11121/4310-7699-0054-1394-CC</a></td>
</tr>
<tr>
<td>Ma, Marshall X</td>
<td><a href="mailto:max7@rpi.edu">max7@rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/4304-5761-5389-3306-CC">http://dx.deepcarbon.net/11121/4304-5761-5389-3306-CC</a></td>
</tr>
<tr>
<td>Fox, Peter</td>
<td><a href="mailto:pfox@cs.rpi.edu">pfox@cs.rpi.edu</a></td>
<td>Rensselaer Polytechnic Institute</td>
<td><a href="http://dx.deepcarbon.net/11121/7699-3078-6331-9782-CC">http://dx.deepcarbon.net/11121/7699-3078-6331-9782-CC</a></td>
</tr>
</tbody>
</table>

## Publications

<table>
<thead>
<tr>
<th>Article</th>
<th>Author</th>
<th>Journal Volume</th>
<th>DCO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance Representation for the National Climate Assessment in the Global Change Information System</td>
<td>Fox, Peter</td>
<td>IEEE Transactions on Geoscience and Remote Sensing</td>
<td><a href="http://dx.deepcarbon.net/11121/6006-7046-2364-5863-CC">http://dx.deepcarbon.net/11121/6006-7046-2364-5863-CC</a></td>
</tr>
</tbody>
</table>

## Awards

<table>
<thead>
<tr>
<th>Award Name</th>
<th>Recipient</th>
<th>Awarding Agency</th>
<th>DCO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science Order of Distinction</td>
<td>Erickson, John</td>
<td>National Center for Interdisciplinary Research</td>
<td><a href="http://dx.deepcarbon.net/11121/6624-4704-2348-7882-CC">http://dx.deepcarbon.net/11121/6624-4704-2348-7882-CC</a></td>
</tr>
</tbody>
</table>

## Grants

<table>
<thead>
<tr>
<th>Grant Recipient</th>
<th>Granting Agency</th>
<th>Award Term</th>
<th>DCO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>West, Patrick</td>
<td>Rensselaer Polytechnic Institute</td>
<td>2009-2012</td>
<td><a href="http://dx.deepcarbon.net/11121/2305-3563-9309-2954-CC">http://dx.deepcarbon.net/11121/2305-3563-9309-2954-CC</a></td>
</tr>
<tr>
<td>Erickson, John</td>
<td>National Center for Interdisciplinary Research</td>
<td>2013-2017</td>
<td><a href="http://dx.deepcarbon.net/11121/6193-6207-7614-8120-CC">http://dx.deepcarbon.net/11121/6193-6207-7614-8120-CC</a></td>
</tr>
</tbody>
</table>

## Fieldwork

<table>
<thead>
<tr>
<th>Fieldwork Title</th>
<th>Location</th>
<th>Leader</th>
<th>Funding Organization</th>
<th>DCO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Collection in East Liboria</td>
<td>East Liboria</td>
<td>Erickson, John</td>
<td>Alfred P. Sloan Foundation</td>
<td><a href="http://dx.deepcarbon.net/11121/6320-7309-1788-7919-CC">http://dx.deepcarbon.net/11121/6320-7309-1788-7919-CC</a></td>
</tr>
</tbody>
</table>
State to date…

- Knowledge network – implements both the collaboration and the integration
- Many means of population
  - User generation
  - Machine generation
- Substantially contributing these enhancements back to open-source community (CKAN, VIVO, GHS)
USGCRP was mandated by Congress in the Global Change Research Act (GCRA) of 1990 (P.L. 101 – 606)

“To provide for development and coordination of a comprehensive and integrated United States Research Program which will assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change.”
• **Coordinates** Federal research to better understand and prepare the nation for global change
• **Prioritizes** and supports cutting edge scientific work in global change
• **Assesses** the state of scientific knowledge and the Nation’s readiness to respond to global change
• **Communicates** research findings to inform, educate, and engage the global community
Global Change Information System (GCIS)

Vision: A unified web based source of authoritative, accessible, usable, and timely information about climate and global change for use by scientists, decision makers, and the public.
Every 4 years


About This Website

This website is intended to make the 2009 National Climate Assessment ("Global Climate Change Impacts in the United States") more accessible to a variety of interested readers. Several features of this site are prototyping greater traceability of source material, better searchability of the assessment as a whole and enhanced access to images and references. For example, in this 2009 report, you will find links to several supporting datasets, including links from clickable images, a more searchable suite of references, a search function for the entire report, and better linking between chapters. Please note however, that this web-based deployment of the 2009 National Climate Assessment is in a trial phase and is intended to help us learn better techniques for future assessments; it is not intended to provide complete, retroactive access to all source material. Lessons learned from this process are being applied in the development of the next synthesis of the National Climate Assessment, due in 2013.
Every 4 years

About This Website

This website is intended to make the 2009 National Climate Assessment ("Global Climate Change Impacts in the United States") more accessible to a variety of interested readers. Several features of this site are prototyping greater traceability of source material, better searchability of the assessment as a whole and enhanced access to images and references. For example, in this 2009 report, you will find links to several supporting datasets, including links from clickable images, a more searchable suite of references, a search function for the entire report, and better linking between chapters. Please note however, that this web-based deployment of the 2009 National Climate Assessment is in a trial phase and is intended to help us learn better techniques for future assessments; it is not intended to provide complete, retroactive access to all source material. Lessons learned from this process are being applied in the development of the next synthesis of the National Climate Assessment, due in 2013.
• “... prepare a summary “traceable account” (a few sentences to a paragraph) that describes the main factors that contributed to the conclusion and level of confidence”

• “In addition to providing a summary traceable account, use the appropriate term below in a parenthetical phrase following the finding to convey to readers the level of confidence associated with the finding”
GCIS approach – uses “linked data”

• Create an **entity** from the structured metadata about each thing – tag with related **concepts**.

• Identify it with a persistent, controlled **identifier**.

• Present with a human readable web page and a **machine interface**.

• Represent all **relationships** between items.
## URI naming

<table>
<thead>
<tr>
<th>Category</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td><a href="http://data.globalchange.gov/person/">http://data.globalchange.gov/person/</a>&lt;unique name&gt;</td>
</tr>
<tr>
<td>Publication</td>
<td><a href="http://data.globalchange.gov/publication/doi/">http://data.globalchange.gov/publication/doi/</a>&lt;doi&gt;</td>
</tr>
<tr>
<td>Project</td>
<td><a href="http://data.globalchange.gov/project/ACMAP">http://data.globalchange.gov/project/ACMAP</a></td>
</tr>
<tr>
<td>Topic</td>
<td><a href="http://data.globalchange.gov/topic/Human-health">http://data.globalchange.gov/topic/Human-health</a></td>
</tr>
<tr>
<td>Image</td>
<td><a href="http://data.globalchange.gov/image/">http://data.globalchange.gov/image/</a>&lt;uuid&gt;</td>
</tr>
<tr>
<td>Figure</td>
<td><a href="http://data.globalchange.gov/report/">http://data.globalchange.gov/report/</a>&lt;reportid&gt;/figure/&lt;figureid&gt;</td>
</tr>
<tr>
<td>Chapter</td>
<td><a href="http://data.globalchange.gov/report/">http://data.globalchange.gov/report/</a>&lt;reportid&gt;/chapter/&lt;chapterid&gt;</td>
</tr>
<tr>
<td>Organization</td>
<td><a href="http://data.globalchange.gov/organization/NASA">http://data.globalchange.gov/organization/NASA</a></td>
</tr>
<tr>
<td>Model</td>
<td><a href="http://data.globalchange.gov/model/">http://data.globalchange.gov/model/</a>&lt;model_name&gt;</td>
</tr>
<tr>
<td>Dataset</td>
<td><a href="http://data.globalchange.gov/dataset/doi/">http://data.globalchange.gov/dataset/doi/</a>&lt;doi&gt;</td>
</tr>
<tr>
<td>Platform</td>
<td><a href="http://data.globalchange.gov/platform/">http://data.globalchange.gov/platform/</a>&lt;platform_name&gt;</td>
</tr>
<tr>
<td>Instrument</td>
<td><a href="http://data.globalchange.gov/instrument/">http://data.globalchange.gov/instrument/</a>&lt;instrument_name&gt;</td>
</tr>
</tbody>
</table>
W3C PROV (starting points..)

Diagram from W3C PROV group and Ivan Herman
# Use Case (UC-1)

<table>
<thead>
<tr>
<th><strong>Name</strong></th>
<th>Discover and visit data center website of dataset used to generate report figure.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>The NCA Report reader sees a figure and wants to know where the data came from.</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>A reader of the NCA is browsing the content via the website. <strong>He/she sees a figure and wants to know where the data came from.</strong> A reference to the publication in which the figure originated appears in the figure caption. Selecting the link to the source publication displays a page of information about the publication including, if available, the publication DOI. The page also includes references to the datasets cited in the publication. Following each of dataset reference links presents a page of information about the dataset, including links back to the agency/data center webpage describing the dataset in more detail and making the actual data available for order or download.</td>
</tr>
<tr>
<td><strong>Actors</strong></td>
<td>Primary Actor - reader of the NCA</td>
</tr>
<tr>
<td><strong>Preconditions</strong></td>
<td>Reader is viewing the NCA online report</td>
</tr>
<tr>
<td><strong>Post Conditions</strong></td>
<td>Reader visits the data center dataset website</td>
</tr>
</tbody>
</table>
| **Normal Flow** | 1) System is presenting the NCA report to the reader in a web site. Presentation includes report figure with caption that includes reference to source publication.  
2) Reader selects publication reference in figure caption  
3) System displays information about publication, including DOI (if available).  
4) Publication information includes publication dataset citations.  
5) Reader selects a dataset cited by the publication.  
6) System displays information about dataset including links to agency / data center webpages where more information and (potentially) data download links are available.  
7) Reader selects the data center link and is redirected to data center dataset webpage. |
### Use Case (UC-1)

**Name**
Discover and visit data center website of dataset used to generate report figure.

**Goal**
The NCA Report reader sees a figure and wants to know where the data came from.

**Summary**
A reader of the NCA is browsing the content via the website. He/she sees a figure and wants to know where the data came from. A reference to the publication in which the figure originated appears in the figure caption. Selecting the link to the source publication displays a page of information about the publication including, if available, the publication DOI. The page also includes references to the datasets cited in the publication. Following each of the dataset reference links presents a page of information about the dataset in more detail and making the actual data available for order.

**Actors**
Primary Actor - reader of the NCA

**Preconditions**
Reader is viewing the NCA online report

**Post Conditions**
Reader visits the data center dataset website

**Normal Flow**
1. System is presenting the NCA report to the reader in a web site. Presentation includes report figure with caption that includes reference to source publication.
2. Reader selects publication reference in figure caption.
3. System displays information about publication, including DOI (if available).
4. Publication information includes publication dataset citations.
5. Reader selects a dataset cited by the publication.
6. System displays information about dataset including links to agency / data center webpages where more information and (potentially) data download links are available.
7. Reader selects the data center link and is redirected to data center dataset webpage.

**Discover and visit data center website of dataset used to generate report figure.**
National Climate Assessment links to GCIS entities

http://nca2009.globalchange.gov/southeast

Southeast

The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity, compared with the rest of the continental United States. The average annual temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, annual average temperature has risen about 2°F, with the greatest seasonal increase in temperature occurring during the winter months. The number of freezing days in the Southeast has declined by four to seven days per year for most of the region since the mid-1970s.

Table of Contents (hide)

1. Rising Temperatures
2. Water Resources
3. Sea-level Rise and Hurricane Intensity
4. Ecosystems
5. Society
6. Adaptation: Reducing Exposure to Flooding and Storm Surge
7. References

Average annual precipitation has increased by 30 percent in the region since 1961. The decline in fall precipitation in South Florida contrasts strongly with the regional average. There has been an increase in heavy downpours in many parts of the region, while the percentage of the region experiencing moderate to severe drought increased over the past three decades. The area of moderate to severe spring and summer drought has increased by 12 percent and 14 percent, respectively, since the mid-1970s. Even in the fall months, when precipitation tended to increase in most of the region, the extent of drought increased by 9 percent.

Climate models project continued warming in all seasons across the Southeast and an increase in the rate of warming through the end of this century. The projected rates of warming are more than double those experienced in the Southeast since 1975, with the greatest temperature increases projected to occur in the summer months. The number of very hot days is projected to rise at a rate greater than the average temperature. Under a higher emissions scenario, average temperatures in the region are projected to rise by about 4.5°F by the 2080s, while a higher emissions scenario yields about 8°F of average warming (with about a 10°F increase in summer temperatures and a much higher heat index). Spring and summer rainfall is projected to decline in South Florida during this century. Except for indications that the amount of rainfall from individual hurricanes will increase, climate models provide divergent results for future precipitation for the remainder of the Southeast. Models project that Gulf Coast states will tend to have less rainfall in winter and spring, compared with the more northern states in the region, because higher temperatures lead to more evaporation of moisture from soils and water loss from plants, the frequency, duration, and intensity of droughts are likely to continue to increase.

http://globalchange.gov/dataset/USHCN_002

The U.S. Historical Climatology Network (USHCN) Version 2 Serial Monthly Dataset


Journal: Bulletin of the AMS

http://globalchange.gov/journal/BAMS

Meene, M.J.
C.N. Williams Jr.
R.S. Vose

http://globalchange.gov/person/235
http://globalchange.gov/person/587
http://globalchange.gov/person/372

Pairwise Homogeneity Adjustment Software

http://globalchange.gov/software/USHCN_V52d.20100217
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity. The temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, an increase in temperature occurring during the winter months. The number of freezing days in the Southeast has decreased, and this trend is expected to continue.

The Southeast is home to a diverse array of ecosystems, including forests, wetlands, and coastal areas. These ecosystems provide important habitat for a wide range of species, including many that are characteristic of the region. The Southeast is also a major agricultural region, with significant production of crops such as cotton, soybeans, and peanuts.

In addition to its natural resources, the Southeast is also home to a rich cultural heritage. The region has a long history of indigenous peoples, and its diverse population includes numerous communities of African, Hispanic, and European descent.

The Southeast is also a major economic region, with a significant presence of manufacturing, agriculture, and service industries. The region is home to numerous cities and towns, including Atlanta, which is the capital of Georgia and a major hub for transportation and commerce.

Overall, the Southeast is a region of great beauty, diversity, and cultural richness, and it continues to play an important role in shaping the future of the United States.
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity. Temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, there has been an increase in temperature occurring during the winter months. The number of freezing days in the Southeast has decreased by 12 days per decade. This trend is consistent with a general warming pattern seen across the northern hemisphere.

*Figure 001: Observed Changes in Precipitation 1901 to 2007*

*Figure 002: Effects of Late Spring Freeze on Crop Yield 1960 to 2000*

*Figure 003: Historicalchanges in Precipitation 1950 to 2000*

*Figure 005: Southern Florida has experienced a nearly 10 percent drop in precipitation in recent decades.*

Image Reference: NOAA/NCDC
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity. Temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, there has been an increase in temperature occurring during the winter months. The number of freezing days in the Southeast has begun to decline.
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity, compared with the rest of the continental United States. The average annual temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, annual average temperature has risen about 2°F, with the greatest seasonal increase in temperature occurring during the winter months. The number of freezing days in the Southeast has declined by four to seven days per year for most of the region since the mid-1970s.
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity, compared...
The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity, compared

While average fall precipitation in the Southeast increased by 30 percent since the early 1900s, summer and winter precipitation declined by nearly 10 percent in the eastern part of the region. Southern Florida has experienced a nearly 10 percent drop in precipitation in spring, summer, and fall. The percentage of the Southeast region in drought has increased over recent decades. Image Reference: NOAA/NODC

Tomas R. Karl

label
Tomas R. Karl

type
Person

works at
USCRP
Traceable accounts...
Key Message & Traceable Account
Key Message vs. “General” Message

 gcis:KeyMessage
  gcis:generalMessage
   gcis:Message
    rdfs:comment
     "Message not directly tied to any one report or traceable account"
   gcis:Message
    rdfs:comment
     "Message tied directly to one report chapter optionally with a traceable account"
Details (don’t blink)
Details (don’t blink)
Interagency Information Integration

GCIS can use relationships between all relevant information about global change across the agencies:

- From observations to datasets to research papers to models to analyses to organizations to people to synthesized reports to human impacts...
- Determine *agency interdependencies* -- An EPA analysis uses a NOAA model dependent on observations from a NASA satellite.
- Can present *unique interagency metrics* "How many papers referenced datasets from a specific satellite?"
- Direct users *back to agency data centers* for more detailed information and the actual content and data.
GCIS Data Mining

Structured information with relationships allows integrated data mining, searching, metrics.

- What projects provided data used to produce figures that were referenced in the 2013 NCA section about coastal sea level rise impacts?
- Which data centers hold data referenced by papers related to forests in the midwest?
- Which agencies have people working on projects related to societal impacts of extreme weather events?
- Show me the latest papers about health impacts of air quality in California. Which datasets were used in the analysis of air quality in California?
State to date …

- NCA3 to be released early 2014
- Fully ‘scraped’ document and GCIS knowledge base to be populated by then
- Planned phase II, capturing traceable account as the content is generated (ask me)…
Thus… progress…

- Integrating – semantics
- Transparent – semantics
- Collaborative – semantics
- Application integration
Thus… progress…

- Integrating – semantics
- Transparent – semantics
- Collaborative – semantics
- Application integration

- But what do I really want?
Means of conduct of research*

- Induction
- Deduction
- Observation
Means of conduct of research*

- Induction
- Deduction

Pattern

Observation
Means of conduct of research*

- Induction
  - Observation
  - Pattern
  - Tentative hyp.
- Deduction
Means of conduct of research

- Induction
  - Observation
  - Pattern
  - Tentative hyp.
  - Theory

- Deduction
Means of conduct of research*

- Induction
  - Observation
  - Pattern
  - Tentative hyp.
  - Theory

- Deduction
  - Theory
Means of conduct of research*

- **Induction**
  - Observation
  - Pattern
  - Tentative hyp.
  - Theory

- **Deduction**
  - Hypothesis
  - Theory
Means of conduct of research*

- Induction
  - Observation
  - Pattern
  - Tentative hyp.
  - Theory

- Deduction
  - Theory
  - Observation
  - Hypothesis
Means of conduct of research*

- **Induction**
  - Observation
  - Pattern
  - Tentative hyp.
  - Theory

- **Deduction**
  - Observation
  - Hypothesis
  - Confirmation
  - Theory
Fundamentally though: putting the SCIENCE back into Data Science

We’ve built capabilities to support induction or deduction and sometimes both, but does this really enable the breadth of science discoveries we seek?

Induction for data science?
So, what about abduction?

No, not the alien kind or the criminal kind...
Huh abduction?

Is a method of logical inference introduced by C. S. Peirce which comes prior to induction and deduction for which the colloquial name is to have a "hunch"
Tell me more Mr. Spielberg

• Leverage our open world (of science), and meaning/semantics, too… oh, and this is all on the web
• Relies on abductive reasoning
  – starts when an inquirer considers
    of a set of seemingly unrelated facts
  – armed with an intuition that they
    are somehow connected and …
  – Is data intensive
  – And this can be a job for
    visualization, among other things…
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.
Abductive platforms – what framework?

• What would this look like in application tools? How to explore ‘hunches’ (hints)?
• If you consent that induction is fundamentally part of how an information system is developed, then how to allow for abduction before induction may be possible?
• And in an open world; collaborative, integrative, transparent -> hyperthesis (cf. hypothesis)
All these generations of mediation are in effect as we conduct research.

From: C. Borgman, 2008, NSF Cyberlearning Report
Thus… progress…

- Integrating – semantics
- Transparent – semantics
- Collaborative – semantics
- Application integration
- Abduction
  – Need models/processes…
  – Ask me: EC$^+$ and WOLCC
Aggregates in First-order Logic

Aggregate expression in FOL:

\[ \text{\#sum}\langle r, px.PartValue(p, t, px, r)\rangle \]

Value of an additive quantity is given by summation of all relevant concurrently active values/effects/changes:

\[
\exists r, px.PartValue(p, t, px, r) \\
\rightarrow \text{Value}(p, t) = \text{\#sum}\langle r, px.PartValue(p, t, px, r)\rangle
\]

But: we have replaced a close summation of 
\[ K \times \text{Value}(\text{Amount}(B), t) \text{ and } K \times \text{Value}(\text{Amount}(C), t) \]
by an open (aggregate) summation in a monotonic logic with open world assumption.
Future Works

Extended Event Calculus:

- Design similar extensions to Event Calculus for other FOL-based formalisms like Fluent and Situation Calculi
- Include durative actions and actions with delayed effects in the extended Event Calculus
- Design abductive reasoning algorithms for continuous-change, and the extended, Event Calculus

Process modeling language:

- Realize a process modeling language for the Semantic Web, with tools to support authoring, parsing, etc.
Thank you

- pfox@cs.rpi.edu
- http://tw.rpi.edu
- @taswegian #twcrpi
• It’s an open world
Semantic Web Standards*

- Ontology - OWL 1.0 (Web Ontology Language, 2004)
- Query - SPARQL 1.0 (SPARQL Protocol and RDF Query Language, 2008)
- OWL 2.0 (2009)
- Taxonomy - SKOS (Simple Knowledge Organization System, 2009)
- Rules - RIF (Rule Interchange Framework, 2010)
- SPARQL 1.1 (in review)
- NB. No service standards!