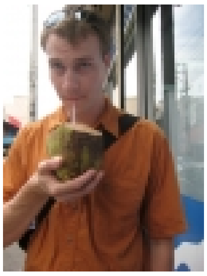
# Accessibility of Earth Systems Science Data **[1]**



Submitted by wheelebk on Tue, 2013-11-19 10:58   Thursday, January 9, 2014 - 13:30 to 15:00
**Event:** Winter Meeting 2014 [2]
**Session Type:** Panel [3]
**Collaboration Area:** Discovery [4]
Documentation [5]
Information Quality [6]
Information Technology and Interoperability [7]
Products and Services [8]
**Abstract/Agenda:**

The current state of affairs in accessing scientific data describing complex earth systems is one that typically involves discrete dataset downloads, but a slowly increasing number of dynamic data services are providing for subsetting or the creation of on-demand derivatives to increase data access and utilization efficiency. There are a growing number of projects in many domains designing and implementing large synthesis data systems to provide access and use of integrated data from many discrete sources. At the same time, there remain data that are essentially inaccessible due to many factors, including lack of digital curation to maintain viable formats, the need for data rescue/preservation, and issues associated with confidentiality or security concerns. One of the aspirations of the government open data initiatives is to enable an increase in activities addressing these issues by government, academia, and private industry that will ultimately establish improved practices in many areas.

This session will include brief presentations from agency staff involved in managing data repositories, producing data syntheses, and providing for the access to earth systems science data followed by an open discussion on challenges and opportunities in data accessibility. The following key questions will be pursued:

- What service and API methods are being employed in the release of scientific data to move beyond the individual dataset download?
- What are the successes and failures in the production of new data synthesis/integration products?
- What are the major impediments in technology, culture, policy, or other areas to improved data accessibility?
- How have the collaborative opportunities brought about through the ESIP community helped to advance capabilities in data accessibility?

**Slides from this session:**
Metadata For Humans and Machines **[9]** from Ted Habermann **[10]**

**Notes:**
Please contact Sky Bristol (sbristol@usgs.gov [11]) and Ben Wheeler (bwheeler@usgs.gov [12]) if interested in presenting/participating in this session.  Organizers are especially interested in garnering a wide range of agency input.

Sky Bristol

· This session will focus on getting access to what metadata describes

· Steve Richards, Ted Habermann – different ways of encoding those access methods, the Sky USGS methods

Ted Habermann, The HDF group

Metadata for Humans and Machines

· Particularly focus on ISO19115 and the ISO family

· Important to think about human and machine readable metadata

· Methods for accessing data has changed

o Data is FTPed to your computer and then you do something to it

· New model

o Have thredds that has more data that provide data via various service providers to users

o This is not really a new idea

· THREDDS = Data + Services

o Important thing about thredds is it allows you to connect data to a service

o Data SErivices – OPenDAP, WCS, NetCDFSubset

o Metadata Services – WMS, NCML, H4MAP

o Allow you to access tools

o Thredds catalogue explains how to do these things

o Want the same data available with multiple services, because we have multiple users and multiple purposes for the data

· nciSO – Serivce metadata

o in NcML

o nciSO added in group

o Thredds makes connection with the services, which is head in the thredds catalogue, not the service itself.

o Extra from cataologue and held in thredds services o   In NcML response to thredds, get a series of little xml elements that describe the services – includes name and url (from thedds)

·         ISO Online Resource

o   Old url… but they became more complex (iso and gcmd responded)

o   Iso added things for humans to understand url

§  This make CI_online resource human readable

o   In iso – there are multiple identification records – can describe a service or data (they are the same)

o   In the revised iso – some tried to get rid of multiple identification records – others made argument that it was too complicated

o   In distribution, there are 2 transfer options – these are complete service descriptions

o   1 of the identification info is data the rest are services

o   in fgdc, there is no service concept

o   order in xml is not important… others assumed in fgdc that it was important

o   service metadata is for machines

o   distribution information is for humans

o   Q (James) – in the last part, links that you don't know what they are to.  Could you address typed links?

§  Service metadata has a title, who runs it; its capability

§  Capabilities have types and parameters

§  Q but that isn't the same as providing the type link

§  But if you convert the iso to rdf – then you might use the service identifiers to get type links

Stephen Richard – Distribution or Service Identification: Two approaches to metadata encoding

·         In metadata record

o   There is a bunch of stuff to keep track of a metadata record

o   Stuff that describes a resource

o   Metadata record is about 1 resource

§  That is what is in the identification record

o   Then there are other things

o   Distribution is how can I get this resource

·         So the record is about the resource (FRBR – functional resource bibliographic record)

·         How do we use metadata

o   Use it to find stuff

o   To evaluate the resource – that is when the other content information comes in

§  Is it likely to be what I need

o   Distribution – is can I get it

·        The Get part

o   Talking about the machine processing of metadata

o   If just urls – that isn’t as much of a problem

o   Particularly – client automation

·        Built a hypermedia driven extension for arcMap

o   Can select something, if map service, it can be added

o   Machine actionable links – search catalogue within arcGIS (or whatever you are using) and the catalogue client can determine if there is a distribution that this software can use

·        CI_OnlineResource

o   Distribution options offered by ISO19139 XML

o   Where do you put these machine actionable links in a metadata record

·        What does the client need

o   Is this resource accessible through a service that I understand

·        Any service will get you a description of what the service can do

o   Get capability, OpenSearch description doc… etc.

o   If software client can use it – it will be aware and can make use of the appropriate document

o   If it is generalized – then it will work well for some things and not for others

·        In the metadata

o   2 places – distribution or service identification

o   Distribution – distributor, format, transfer options

o   ServiceID – duplicate dataID stuff and data offerings…

o   Don’t need to know everything, just the “get” capabilities

·        If in service section… look service type and url… essentially CI_OnlineResource

·        If in distribution section – linkage/url and protocol – still in CI_OnlineResource

o   Is there 2 places or 1 place to look for resource

o   Makes sense to put it in 1 place – distribution section

· Is it about all these different things or just the dataset

· If there parameters that you need to know about for a request

o If use service, put it in the text description – if it isn't obvious in capability document

· Metadata record about a dataset should have dataIdentification, access dataset via different distributions, service information is for documenting a service

· Would like to have 1 way to do things

· Q – do you have an opinion on the serf (service entry resource format) format, has interaction of services but it isn't the main point?

o Describes actual software tools – resource is actually an application

o ?

· Q – James – have you thought about using RDF and RDFs to solve these problems

o In stead of XML they are attributes

o In RDF, how do you assign – why make them play a guessing game

· Q – Ted – there are way to put this information in that involves shoe horning or not. All services have distribution documents – future may have more complexity. But passing of parameters, chaining, if using bpml prarmeters – as we get to more complex pictures the serviceinformation is more important, but when you need to grow/innovate then have more problems

o Here is the services and then there is a service metadata record

o Ted – more difficult to architect/explain a link of multiple records

· Q – what is the lifetime of the metatdata that you put out here – what happens if urls are no longer valid

o Ubiquitous problem – service needs lifetime

o It is a management question – if maintenance a system then responsible to keep it work

o Even if get a machine readable list of services… but still need to get the correct url

o Ted – need a mechanism to test the links – link rot

Sky Britol – USGS ScienceBase

· ScienceBase is a digital repository and data management

o Similar to ECHO and GStore

o User story driven process

· Data File Inputs

o Will take any kind of file

o Certain kinds of files have certain things done with them when they are in the provided

- Metadata(Item) Inputs

o An item is an item is an item… anything is an item

o Still dealing with FGDC and other formats

o Have different types of harvesters

- Technology

o Inbound API

o MetaBase (MongoDB)

o Complex geometry store (PostGIS)

- Service

o Metadata Services and Data Services

- Primary output interfaces

o Html,

o sbJSON,

o ISO19139,

o ATOM, MODS, CSV, CSDGM

- Downstream uses

o Drupal Modules

o Python and R

o JQuery and Similar Frontends

o ArcGIS Desktop

o CSW Search

o OAI-PMH Harvest – feeds data.gov

- Also have an about page and data documentation page

- Government agencies are working to feed data.gov

o USGS data is coming from various sources

o Problem – read some metadata document.. which do I read… how do I deal with it

o Might do a great encoding job in an ISO record – but it won't necessary read it correctly (machine vs. human readable sections)

- ScienceBase examples

- Have service ID information duplicated some of them in distribution information

· This problem tracks from getting data into the repository, getting at it, serving it up to repository, and then downstream users – decisions are important

· Have cases using the “Steve” method, but not currently the “Ted” method

· Q (Ted) – this is nice – how these things are together in a system that services a lot for users. Plea for real world examples, need these to figure out how we can move forward. So, you are using the Ted method

· Q – uploading a dataset, how many are you currently keeping track of

o 3.5 million. Demo upload is 1 item

- Q - how does it relate to archieve

o trying to provide of all data. user want specific record, is there a haresting method for other areas

- Q - are you trying to look for tracability back through sciencebase

o yes, ex. entire historic topomap is houses in sciencebase - 2.5 million of those products

**Session Leads:**

**Name:** Sky Bristol [13]
**Organization(s):** U.S. Geological Survey [14]

**Presenters:**

**Name:** Sky Bristol [13]
**Organization(s):** U.S. Geological Survey [14]

**Name:** Ted Habermann [15]
**Organization(s):** The HDF Group [16]
**Email:** thabermann@hdfgroup.org [17]

**Name:** Stephen Richard [18]
**Organization(s):** Arizona Geologic Survey [19]
**Email:** steve.richard@azgs.az.gov [20]

**Notes takers:**

**Name:** Kelly Monteleone [21]
**Organization(s):** University of New Mexico [22]
**Email:** krbm@unm.edu [23]

**Source URL:** http://commons.esipfed.org/node/1845

**Links:**
[1] http://commons.esipfed.org/node/1845
[2] http://commons.esipfed.org/taxonomy/term/1029
[3] http://commons.esipfed.org/session-type/panels
[4] http://commons.esipfed.org/collaboration-area/discovery
[5] http://commons.esipfed.org/collaboration-area/documentation
[6] http://commons.esipfed.org/collaboration-area/information-quality

[7] http://commons.esipfed.org/collaboration-area/information-technology-and-interoperability
[8] http://commons.esipfed.org/collaboration-area/products-and-services
[9] https://www.slideshare.net/tedhabermann/metadata-for-humans-and-machines
[10] http://www.slideshare.net/tedhabermann
[11] https://mail.google.com/mail/?view=cm&amp;fs=1&amp;tf=1&amp;to=sbristol@usgs.gov
[12] https://mail.google.com/mail/?view=cm&amp;fs=1&amp;tf=1&amp;to=bwheeler@usgs.gov
[13] http://commons.esipfed.org/node/323
[14] http://commons.esipfed.org/taxonomy/term/397
[15] http://commons.esipfed.org/node/1847
[16] http://commons.esipfed.org/taxonomy/term/822
[17] mailto:thabermann@hdfgroup.org
[18] http://commons.esipfed.org/node/1706
[19] http://commons.esipfed.org/taxonomy/term/312
[20] mailto:steve.richard@azgs.az.gov
[21] http://commons.esipfed.org/node/430
[22] http://commons.esipfed.org/taxonomy/term/297
[23] mailto:krbm@unm.edu