DataONE Architecture and How to Become a Member Node [1]

Submitted by superadmin on Fri, 2012-06-29 12:58   Tuesday, July 17, 2012 - 13:30 to 15:00
Tuesday, July 17, 2012 - 15:30 to 17:00
Event: Summer Meeting 2012 [2]
Session Type: Workshop [3]
Media/Video: session recording (streaming) [4]
session recording (download) [5]
Expertise Level: Beginner [6]
Identifier: doi:10.7269/P3057CV6
Collaboration Area: Information Technology and Interoperability [7]
Abstract/Agenda:

This session will provide an overview of the DataONE architecture, explain the benefits for groups and institutions to publish data as a Member Node, provide hands-on tutorials in how to establish a Member Node using various software systems, and demonstrate how to use the DataONE web services to access content from client applications. By the end of the session, participants will have a detailed knowledge of the design of the DataONE architecture, the services that DataONE provides to its Member Nodes, and the technical details needed to establish a Member Node at their own site or institution and to build client applications. The session will target information managers, graduate students, post-docs, faculty, and research technicians who manage environmental data at a site as well as people who develop data management and analysis software supporting researchers.

Notes:
(Collaborative notes from speakers)

DataONE Architecture and How to Become a Member Node

Room: Pyle Center 335 Expertise Level: Beginner

Collaboration Area: Information Technology and Interoperability

Teaser: Overview of the @DataONEorg architecture for info mgrs, grad students, post-docs, faculty, research techs who manage env data

Agenda

Tue, 07/17/2012 - 13:30 to 15:00

Overview of DataONE (Rebecca Koskela, UNM) (20 mins, 10 mins questions)

- terminology (CN, MN, ITK)

- current MNs (chart)

Becoming a Member Node (Matt Jones, UCSB) (40 minutes, 20 minutes discussion)

Tue, 07/17/2012 - 15:30 to 17:00

Identifiers and Packaging in DataONE (Matt Jones) (15/15)

Investigator Toolkit

DataUp Excel plugin (Carly Strasser) (15/15)

[Tools Overview (Matt Jones) (15/15)]

Notes

Questions:

What are the criteria for becoming a Member Node?

Is there a difference between EarthCube and DataONE?

There are some overlaps with DataONE - but definitely should be interoperability between the two. D1 playing a technical role in EarthCube.

How long is D1 funded? 5 years, then reapply for another 5 years

What cyberinfrastructure has D1 built?

The overlay for the MNs to become MNs; CN software; and tools in the ITK

Skipping to "How To:" section

What type of data? We define environmental data fairly broadly - reference to tag line

Doesn't have to be geo-located

Participants:

UAF- GI & ASF plus the ALISON data (ice data collected by school children)

Arizona

NGDS - Alabama

Are you collecting information about who is accessing the data?

Log data about access to datasets - although some MNs allow anonymous access so would just know that someone accessed the dataset - otherwise would know who accesssed it

How is Merritt different that DSpace or Fedora?

How many people are using system?

How scalable will the system be? Could it handle a 4 PT dataset?

Identifiers

ESIP did a survey and reccomend:

DUI for datasets

UIDs for granules

If a replication target, then need to be able to accept other MNs identifier system

Why not URL or URI? URI can typically be broken - change of server name will break the

URL or URI. Is the resolution service reliable?

Social process of persistence not a technical process

Does D1 count the citations that a dataset receives? Don't track citations ourselves but

groups like Total Impact are trying to mainstream dataset citation

Packing

Lack of agreement of what is a dataset

Challenging when talking about data citation

Package is the means for someone to describe what they mean by dataset

In EZID, establishing similar idea of packaging

DataCite has created a new metadata standard with new namespace

Similar to decisions made by NOAA - archive around a packaging concept but have a catalog file describe what's in a package

Sensor streams can cause a problem for some groups because don't want to partition the data

USGS - Water: many projects that generate data that need to go into a institutional repository

Data coming into a database in real time - how should this work? Discrete snapshots seem

to work - then can assign a unique identifier

D1 might use DOIs - are investigating this now - looking at assigning identifiers to smaller granules. Advice to read the ESIP summary

Actually talking about fine-grained provenance - do you need to have an identifier to define the provenance or is there sufficient information available to identify the granularity?

How to follow data derivations is being worked on by the D1 Provenance & Workflow WG

Implementation is sticking with the coarse grain identification for right now.

Annotation process is closely related to this.

At what level can data be self-describable? Depends of the metadata that accompanies the data

Will the results of Carly's survey on spreadsheets be shared? Yes, going to publish and some results already on blog

Model of maximum scientist involvement - what likes best about DataUp

Browse image is important part of the package - especially for educators - Figshare has a preview of the data

Powerful for discovery

Question about choice of C# for programming language of DataUp - was created by Microsoft - hoping to find a summer student to translate to another programming language

First thing is to get DataUp to communicate with D1 API so then would be able to use any Tier 3 Member Node

NOTES by Kelly Monteleone:

Overview of Data One – by Rebecca Koskela

- NSF funded dataNet project – one of the first, August will be 4th year
- Based on book – 4th Paradigm – increase in data intensive research
- NSF wants funded projects' data available
- Top of pyramid = intensive science
- 80:20
  - 80% formatting and 20% analyzing
  - Long tail – specialized repositories vs orphaned dated
- 3 goals
  1. Build on cyberinfrastructure (ex. LTER)
  2. Create new cyberinfrastructure
  3. New communities of practice
- 2 equal parts 1) cyberinfrastrcuture 2) community engagement

- Cyberinfastructure
  1. Member node – existing repositories
  2. Coordinating nodes – connect repositories – copy metadata NOT data
     - 3 Act as replicatory node – UC Santa Barbara, UNM< Oakridge Tenn Campus
  3. Investigators Toolkit ex. Matlab, Mendeley
- Federation à want members across the world
  - Public release on Thursday July 19, 2012 (2 days)
  - Have 7 nodes – divsrity in tys of communities they serve (6 in US)
- Current tools – python & Java, DMPTool, ONEMercury
- Coming soon (tools) – ONE-R, DAtaUp, ONEDrive (butterfly = morpho à metadata)
- ONEShare – repository at New Mexico – excel and will work with dataUP
- Looked at the data cycle – DataONE has tools for each part of the cycle
- DMPTool à used by many universities for data plans required for funding
- VT = viztrails, discussion with matlab
- ONEMercury – text or sptial/map search
  - Able to filter results by multiple filters
  - Has "data available field" - # based on how member node is set-up
  - Returns – unique ID, file types, size, download button
  - Can save results in Zortero (lost visual for webx)
- Example – bird migration and green-up time
  - Linked with terragrid for better processing
- Create Educational Models & animations – free for download

- Large team = US (mainly), Europe, Australia

Questions

- What criteria are there for becoming a member node or where do you find the information
    - Various requirements, 1st require interest
    - Information on website (and a place to ask questions) under member node
    - Has to have environmental data, want to share, and install software
        - Greg Ederer from NASA Goddard – MODIS data producer – looking how make availabe
- List of pending node?
    - Next presentation

DataONE Member Nodes – Matt Jones

Questions

- How does DataONE relate to EarthCUBE – what difference especially considering both looking at 80:20 problem
    - EarthCUBE is broader – covers governance in science life cycle but they do overlap
    - EarthCube is newer (starting interoperability discussion)
    - DataONE wants to fit into EarthCUBE
    - Need to work as a community
- What is the outlook for DataONE funding
    - Have 5 yr of funding + 2 yr option, then can apply for 5 more years at lower rate
    - Need sustainable portion from NSF à require sustainability plan
    - Has sustainably and governance working group but need different models for different parts of the organization
- How is dataONE creating cyberinfastrucure
    - Member nodes already have infrastructure à building venure between nodes & tools/client tools to talk to nodes
-

Actions:
Beta test for DataUP starting July 17, final July 25, Public Sept 4

to participate in beta test email: carlystrasser@gmail.com [8]

Notes takers:

Name: Kelly Monteleone [9]
Organization(s): University of New Mexico [10]
Email: krbm@unm.edu [11]

Teaser: Overview of the @DataONEorg architecture for info mgrs, grad students, post-docs, faculty, research techs who manage env data
Keywords: DataOne [12]

Source URL: https://commons.esipfed.org/node/435

Links
[1] https://commons.esipfed.org/node/435
[2] https://commons.esipfed.org/event/summer-meeting-2012
[3] https://commons.esipfed.org/session-type/workshop

[4] https://esipfed.webex.com/esipfed/ldr.php
[5] https://esipfed.webex.com/esipfed/lsr.php
[6] https://commons.esipfed.org/taxonomy/term/260
[7] https://commons.esipfed.org/collaboration-area/information-technology-and-interoperability
[8] mailto:carlystrasser@gmail.com
[9] https://commons.esipfed.org/node/430
[10] https://commons.esipfed.org/taxonomy/term/297
[11] mailto:krbm@unm.edu
[12] https://commons.esipfed.org/taxonomy/term/299