Member Nodes - MNDeployment #7865

Hydroshare (CUAHSI)

2016-08-15 17:56 - Laura Moyers

Status:	Planning	**Start date:**	2019-08-05
Priority:	Normal	**Due date:**	
Assignee:	John Evans	**% Done:**	10%
Category:		**Estimated time:**	0.00 hour
Target version:			
Latitude:		**MN_Date_Online:**	
Longitude:		**Name:**	
MN Description:		**Logo URL:**	
Base URL:		**Date Upcoming:**	
NodeIdentifier:		**Date Deprecated:**	
MN Tier:		**Information URL:**	
Software stack:		**Version:**	

Description

CUAHSI - the Consortium of Universities for the Advancement of Hydrologic Science

https://www.cuahsi.org/

Dave and Jeff Horsburgh, jeff.horsburgh@usu.edu are in contact.

MN Descripton Worksheet

Subtasks:

Task # 8830: Requirements Analysis **New**

History

#1 - 2016-08-29 18:42 - Dave Vieglais

Message sent to Jeff et al 2016-08-01:

There's a couple aspects of the current DataONE infrastructure that CUAHSI could leverage to facilitate participation with DataONE.

Mutability is kind of supported through the use of Series Identifiers, which are simply persistent identifiers that always refer to the most current version of a series of data set revisions. Specific versions of a dataset are still immutable however. The main benefit of SIDs is that one can make a reference to the most recent version of a dataset rather than just a specific version.

Service Registration may be more applicable to the CUAHSI model. Service Registration is basically the description of a service by a metadata document. That metadata includes reference to the type and location of services offered by a Member Node (the service may or may not be co-located with the Member Node).

The basic concept is that if you were offering a service like a WFS or SOS, or perhaps hydroshare, that service is described by a metadata document (we are using ISO19119 semantics), and then discovered much like other data through the search interface.

The disadvantage of service registration is that only the metadata is preserved by DataONE - the service and content accessible through that are managed entirely by the service provider.

That said, the features of hydroshare that you mention seem like a good fit with the version 2 capabilities of DataONE and should make setting up a Member Node service much more tractable. It would be good to schedule a time for discussion in the not too distant future to identify a plan for moving forward.

#2 - 2018-01-25 17:49 - Amy Forrester

ePad Notes Consolidation
Dave and Jeff Horsburgh corresponding
10/9/17 - Dave will contact Jeff again
10/16/2017 - still waiting for a message back from Jeff

#3 - 2019-07-23 00:14 - Matthew Jones

Matt J spoke with Jeff Horsburgh (jeff.horsburgh@usu.edu) at a data repositories workshop in Spring 2019, and then with Martin Seul (MSeul@cuahsi.org) at the 2019 summer ESIP. They both expressed interest in having CUASHI join DataONE using their HydroShare collection, which is found at https://hydroshare.org. Each entry in HydroShare is assigned a resource identifier, and is associated with a DOI, and has a schema.org entry in their landing page. They maintain a sitemap at /sitemap.xml. Here's an example SO entry from one of their landing pages:


```
{
    "@context": {
        "@vocab": "http://schema.org/",
        "geolink": "http://schema.geolink.org/1.0/base/main#"
    },
    "@type": "Dataset",
    "additionalType": ["http://schema.geolink.org/1.0/base/main#Dataset", "http://vivoweb.org/ontology/core#Da
taset"],
    "name": "R script to obtain continuous daily estimates of whole-stream metabolism",
    "description": "Generating daily estimates of WSM. We show the procedure to generate daily estimates of on
e station whole\u002Dstream metabolism using water, reach and weather data for the period of interest. Read th
e \u0022Supplemental_Text_S4.pdf\u0022 for detailed description of the Daily.Metabolism.R script.\u000D\u000A\
u000D\u000ADOI: 10.6084/m9.figshare.3422272\u000D\u000A\u000D\u000ANote: the \u0022AAAA\u002DMM_inputs.csv\u00
22 files are the monthly dates files used for the estimation of WSM for the period of August 2010 through Dece
mber 2014. To run the Daily.Metabolism.R script only one of these files is required.\u000D\u000A",
    "url": "https://www.hydroshare.org/resource/aadd7dd60f31498590de32c9b14446c3/",
    "version": "2017-06-04",
     "isAccessibleForFree": true,
    "keywords": "[\u0022Whole\u002Dstream metabolism\u0022, \u0022Time series\u0022, \u0022River ecohydrology\
u0022]",

        "license": "http://creativecommons.org/licenses/by/4.0/",


    "citation": "Villamizar, S., H. Pai (2016). R script to obtain continuous daily estimates of whole-stream
metabolism, HydroShare, http://www.hydroshare.org/resource/aadd7dd60f31498590de32c9b14446c3",
    "includedInDataCatalog": {
        "@id": "https://www.hydroshare.org"
    },
    "distribution": {
        "@type": "DataDownload",
        "contentUrl": "https://www.hydroshare.org/hsapi/resource/aadd7dd60f31498590de32c9b14446c3/",
        "encodingFormat": "application/zip"
    },


    "spatialCoverage": {
        "@type": "Place",
        "geo": {

            "@type": "GeoShape",
            "box": ",  , "

        }
    },

    "creator": [

        {
            "@id": "/user/418/",
            "@type": "Role",
            "additionalType": "http://schema.geolink.org/1.0/base/main#Participant",
            "roleName": "Author",
            "url": "/user/418/",
            "creator": {
                "@id": "/user/418/",
                "@type": "Person",
                "additionalType": "http://schema.geolink.org/1.0/base/main#Person",
```

```
                "name": "Sandra Villamizar",
                "url": "/user/418//"
            }
        }
        ,

        {
            "@id": "",
            "@type": "Role",
            "additionalType": "http://schema.geolink.org/1.0/base/main#Participant",
            "roleName": "Author",
            "url": "",
            "creator": {
                "@id": "",
                "@type": "Person",
                "additionalType": "http://schema.geolink.org/1.0/base/main#Person",
                "name": "Henry Pai",
                "url": "/"
            }
        }


    ],
    "provider": {
        "@id": "https://www.hydroshare.org",
        "@type": "Organization",
        "additionalType": "http://schema.geolink.org/1.0/base/main#Organization",
        "legalName": "HydroShare",
        "name": "HydroShare",
        "url": "https://www.hydroshare.org"
    },
    "publisher": {
        "@id": "https://www.hydroshare.org"
    }
}
```


Issues I immediately noticed that we may want to ask about:

- access to the landing page seemed to require a login
- the SO entry is missing the DOI identifier
- the SO entry does not map between their resource ID and their DOI to indicate they are the same resource
- people are identified with local user identifiers (/user/401) rather than e.g. ORCIDs
- the SO entry appears to not link to more detailed ISO or other metadata

#4 - 2019-07-23 17:25 - Amy Forrester

- Status changed from New to In Review

- % Done changed from 0 to 80

#5 - 2019-07-24 20:20 - Amy Forrester

Call scheduled for 8/1/19 during HydroShare all hands meeting

Meeting notes: CUAHSI Discovery Worksheet

#6 - 2019-08-01 18:01 - Amy Forrester

- Description updated

- Assignee changed from Laura Moyers to Amy Forrester

- Subject changed from CUAHSI - Consortium of Universities for the Advancement of Hydrologic Science to Hydroshare (Cuahsi)

#7 - 2019-08-01 18:12 - Amy Forrester

- Subject changed from Hydroshare (Cuahsi) to Hydroshare (CUAHSI)

#8 - 2019-08-01 18:14 - Amy Forrester

- Description updated

#9 - 2019-08-27 14:43 - John Evans

- File resourcemetadata.xml added

Some issues with CUAHSI thusfar:

The metadata documents have some custom elements (namespace "hsterms") that is specific to hydroshare. Would this then require a custom format ID?

The metadata validation actually errors even before that, though, complaining about the rdf namespace

```
Element '{http://www.w3.org/1999/02/22-rdf-syntax-ns#}RDF': No matching global declaration available for the v
alidation root.
```

Also, as noted above, the DOI isn't where we expect to find it, i.e. not in the @id field.

#10 - 2019-08-27 16:39 - Amy Forrester

- Assignee changed from Amy Forrester to John Evans

- Status changed from In Review to Planning

- % Done changed from 80 to 10

#11 - 2019-09-03 19:32 - John Evans

If we run the site checker on the Hydroshare/CUAHSI top-level sitemap, we see the following

```
$ d1-check-site https://www.hydroshare.org/sitemap.xml
2019-09-03 14:54:45,184 - datatone - INFO - Requesting sitemap document from https://www.hydroshare.org/sitema
p.xml
2019-09-03 14:54:49,444 - datatone - INFO - Requesting sitemap document from https://www.hydroshare.org/sitema
p-pages.xml
2019-09-03 14:54:49,570 - datatone - INFO - Extracted 3 from the sitemap document.
2019-09-03 14:54:49,570 - datatone - INFO - 0 records skipped due to lastmod time.
2019-09-03 14:54:49,570 - datatone - INFO - Looking to process 3 records...
2019-09-03 14:54:49,571 - datatone - INFO - Requesting https://www.hydroshare.org/...
2019-09-03 14:54:49,752 - datatone - WARNING - Skipping https://www.hydroshare.org/ due to "SkipError('Could n
ot locate a JSON-LD <SCRIPT> element with @type "Dataset".')".
2019-09-03 14:54:49,752 - datatone - INFO - Requesting https://www.hydroshare.org/accounts/login/...
2019-09-03 14:54:49,940 - datatone - WARNING - Skipping https://www.hydroshare.org/accounts/login/ due to "Ski
pError('No JSON-LD <SCRIPT> elements were located.')".
2019-09-03 14:54:49,940 - datatone - INFO - Requesting https://www.hydroshare.org/apps/...
2019-09-03 14:54:50,594 - datatone - WARNING - Skipping https://www.hydroshare.org/apps/ due to "SkipError('No
 JSON-LD <SCRIPT> elements were located.')".
2019-09-03 14:54:50,595 - datatone - INFO - Requesting sitemap document from https://www.hydroshare.org/sitema
p-resources.xml
2019-09-03 14:54:56,606 - datatone - INFO - Extracted 7827 from the sitemap document.
2019-09-03 14:54:56,608 - datatone - INFO - 0 records skipped due to lastmod time.
2019-09-03 14:54:56,609 - datatone - INFO - Looking to process 7827 records...
2019-09-03 14:54:56,616 - datatone - INFO - Requesting https://www.hydroshare.org/resource/120eac90fa3b45bab47
3e1914f686ca0/...
2019-09-03 14:54:57,890 - datatone - ERROR - Unable to process https://www.hydroshare.org/resource/120eac90fa3
b45bab473e1914f686ca0/ due to "RuntimeError('JSON-LD missing top-level "@id" key.')".
2019-09-03 14:54:57,890 - datatone - INFO - Requesting https://www.hydroshare.org/resource/aadd7dd60f31498590d
e32c9b14446c3/...
2019-09-03 14:54:59,217 - datatone - ERROR - Unable to process https://www.hydroshare.org/resource/aadd7dd60f3
1498590de32c9b14446c3/ due to "RuntimeError('JSON-LD missing top-level "@id" key.')".
2019-09-03 14:54:59,218 - datatone - INFO - Requesting https://www.hydroshare.org/resource/cf575f78146a4c7b88e
da3f3463f6555/...
2019-09-03 14:54:59,461 - datatone - WARNING - Skipping https://www.hydroshare.org/resource/cf575f78146a4c7b88
eda3f3463f6555/ due to "SkipError('No JSON-LD <SCRIPT> elements were located.')".
2019-09-03 14:54:59,461 - datatone - INFO - Requesting https://www.hydroshare.org/resource/4ab0be5da4b248b48d2
507ac689ccd9a/...
2019-09-03 14:54:59,703 - datatone - WARNING - Skipping https://www.hydroshare.org/resource/4ab0be5da4b248b48d
2507ac689ccd9a/ due to "SkipError('No JSON-LD <SCRIPT> elements were located.')".
2019-09-03 14:54:59,703 - datatone - INFO - Requesting https://www.hydroshare.org/resource/2d35730a34c94d91b17
f9929807fbf4f/...
2019-09-03 14:55:01,028 - datatone - ERROR - Unable to process https://www.hydroshare.org/resource/2d35730a34c
94d91b17f9929807fbf4f/ due to "RuntimeError('JSON-LD missing top-level "@id" key.')".
2019-09-03 14:55:01,028 - datatone - WARNING - Error threshold reached.
2019-09-03 14:55:01,029 - datatone - INFO - Shutting down...
2019-09-03 14:55:01,029 - datatone - INFO - Cancelling 1 outstanding tasks.
2019-09-03 14:55:01,029 - datatone - ERROR - CancelledError()
2019-09-03 14:55:01,029 - datatone - INFO - Successfully processed 0 records.
```

What's going on is that Hydroshare has nested sitemaps. The top-level sitemap points to two nested documents:

1. "/sitemap-pages.xml" - This sitemap terminates with 3 landing pages, in none of which are we are interested. They actually cause errors because none of the 3 landing pages have an SO element. Maybe we should just skip such landing pages without counting it as an error?
2. "/sitemap-resources.xml" - This is the sitemap for what we want. In fact, we could just specify this URL directly to avoid the spurious landing pages.

When we finally start hitting the landing pages with SO elements, we error out immediately due to the lack of an @id element at the top level, in which we expect to find a DOI that looks like a URI. For datasets that are PUBLISHED (this is specified in the landing page HTML), there is an top-level "identifier" key that does have the DOI. Here's an example of the SO in such a case:

```
    "identifier": {
        "@type": "PropertyValue",
        "additionalType": ["http://schema.geolink.org/1.0/base/main#Identifier", "http://purl.org/spar/datacit
e/Identifier"],
        "propertyID": "http://purl.org/spar/datacite/doi",
        "url": "https://doi.org/10.4211/hs.1396774b293144689a66080739da8f44",
        "value": "10.4211/hs.1396774b293144689a66080739da8f44"
    },
```

So instead of '@id', they seem to use "identifier".

Even if the @id were present, the validation would fail upon going further because there is no top-level 'encoding' map, and therefore no 'contentUrl', 'description', 'dateModified', or 'encodingFormat', because they are expected inside the encoding map.

Hydroshare does provide a 'distribution' map with a 'contentUrl' key. However, this just points to the landing page, which we already have. The metadata document is not directly referenced; we have to construct a URL for the zipped baggit archive

There is a top-level "version" key that basically reflects the same idea as 'lastModified'. The hydroshare sitemaps do not include "lastModified" elements.

There is a top-level "description" key that serves the same purpose as the one we expect in the 'encoding' map.

#12 - 2019-09-03 19:40 - John Evans

Since Hydroshare has the metadata document in the baggit zip archive, whose URL we have to construct, it seems that Hydroshare can't really meet our recommendations for SO. But all the pieces are there (just not where we expect them).

#13 - 2019-09-09 19:46 - John Evans

I've constructed an XSL stylesheet that transforms the hydroshare metadata documents into something that validates against http://ns.dataone.org/metadata/schema/onedcx/v1.0.

Using a slendernodes setup, I can successfully harvest about 175 out of about 235 PUBLISHED documents from their sitemap setup when the stylesheet is employed. For those 60 failures, they are mostly a mixture of I/O errors in the asyncio/aiohttp layers that maybe can be addressed through different timeout or chunking setups. Unsure about that. They sometimes resolve upon retries.

There are at least two cases where the landing page issues a pop-up dialogue asking the user to say yes to a wikipedia-style license agreement, and at least one case of a bad zip file on hydroshare's end.

#14 - 2019-09-17 14:33 - John Evans

- File simple_d1_dublincore.xsl added

Style sheet here

Files

resourcemetadata.xml	6.94 KB	2019-08-27	John Evans
simple_d1_dublincore.xsl	8.08 KB	2019-09-17	John Evans