Member Nodes - MNDeployment #3230

ARM - Atmospheric Radiation Measurement member node

2012-09-07 00:21 - Dave Vieglais

Status:	Planning	**Start date:**	2013-10-01
Priority:	Normal	**Due date:**	
Assignee:	John Evans	**% Done:**	10%
Category:		**Estimated time:**	0.00 hour
Target version:	Deploy by end of Y4Q1		
Latitude:	35.93	**MN_Date_Online:**	
Longitude:	-84.30	**Name:**	ARM - Atmospheric Radiation Measurement Research Facility
MN Description:	ARM focuses on obtaining continuous measurements—supplemented by field campaigns—and providing data products that promote the advancement of climate models. Serving users worldwide, the ARM Data Center collects and archives approximately 25 terabytes of data per month. ARM data include routine data products, value-added products (VAPs), field campaign data, complementary external data products from collaborating programs, and data contributed by ARM principal investigators for use by the scientific community. Data quality reports, graphical displays of data availability/quality, and data plots are also available from the ARM Data Center.	**Logo URL:**	https://raw.githubusercontent.com/DataONEorg/member-node-info/master/production/graphics/web/ARM.png
Base URL:		**Date Upcoming:**	2019-11-25
NodeIdentifier:	urn:node:ARM	**Date Deprecated:**	
MN Tier:	Tier 1	**Information URL:**	https://www.archive.arm.gov/
Software stack:	GMN	**Version:**	

Description

This issue captures the activity associated with deployment or otherwise of the "Atmospheric Radiation Measurement" member node.

Subtasks:

Task # 4015: (Waiting For) Resolution of mutability issue into production (ARM Archive)	**Closed**
Story # 8589: ARM: Re-Discovery & Planning	**New**
Task # 8590: ARM: Requirements Analysis	**New**
MNDeployment # 8825: Create test certificate	**Closed**

History

#1 - 2013-01-30 20:59 - John Cobb

- translation missing: en.field_remaining_hours set to 0.0

- Subject changed from Deploy the ARM - Atmospheric Radiation Measurement member node to ARM - Atmospheric Radiation Measurement member node

- Start date deleted (2012-09-07)

#2 - 2013-08-05 23:05 - Laura Moyers

- Latitude set to 35.93

- Longitude set to -84.30

#3 - 2013-08-14 14:46 - Bruce Wilson

- Assignee set to Bruce Wilson

#4 - 2013-08-23 18:10 - Bruce Wilson

per discussion 2013-08-23 with Bob Cook, this MN depends both on dealing with mutable science metadata as well as dealing with streaming data. A good start would be getting the showcase datasets into DataONE. That requires making mercury MN a more drop-in process in the face of mutability.

#5 - 2013-10-09 16:39 - Bruce Wilson

- Software stack set to Custom

- OAI-PMH Stack set to Custom, based on OAICat

#6 - 2014-05-19 12:19 - Bruce Wilson

- Target version set to 317

- Due date set to 2019-07-31

- Start date set to 2014-05-19

#7 - 2014-05-19 12:27 - Bruce Wilson

- Due date deleted (2019-07-31)

- Start date deleted (2014-05-19)

- Target version deleted (317)

#8 - 2016-02-22 22:45 - Laura Moyers

- Status changed from New to Deferred

Deferred indefinitely (see MNW notes 2/22/16)

#10 - 2017-07-20 19:34 - Laura Moyers

- Assignee changed from Bruce Wilson to Laura Moyers

At the 7/20/17 MN Forum, Aaron Stokes said that ARM is considering setting up a DataONE Member Node to increase the visibility and discoverability of their data. It is likely that a Tier 1 MN would serve their needs, but Aaron is going to inquire if they will need any authentication for access or if writing to the ARM MN is desired. I think this is unlikely.

Laura to look at current ARM data portal https://www.arm.gov/data to gather more information about the types, quantities, etc. of content that ARM currently serves up.

Some questions to be addressed: will ARM wish to provide snapshots of the streaming content, described by science metadata? will the data to be exposed be "updated" daily, requiring daily obsolescence? will the data be by site, by sensor, by site-x-sensor, etc? will the ARM MN be exposing science metadata only with links back to the ARM repository for download of science data?

The technical solution for an ARM MN could be similar to SDC and other Mercury-based MNs, but this might be a good opportunity to explore a GMN solution.

#11 - 2017-07-20 19:34 - Laura Moyers

- % Done changed from 100 to 10

- Status changed from Deferred to Planning

#12 - 2017-08-24 02:00 - Laura Moyers

- Assignee changed from Laura Moyers to Monica Ihli

- OAI-PMH Stack deleted (Custom, based on OAICat)

- Software stack changed from Custom to GMN

- MN Tier set to Tier 1

- NodeIdentifier set to urn:node:ARM

Early conversations indicated the ARM MN would be a custom implementation based on an OAI-PMH stack based on OAICat. Recent conversations (summer 2017) suggest that GMN might be a more appropriate MN software solution.

Meeting planned for 1pET Thursday 24 August with Aaron and Ranjeet to discuss how a GMN implementation might work, etc.

#13 - 2017-08-24 02:01 - Laura Moyers

- Target version set to Deploy by end of Y4Q1

#14 - 2017-08-24 21:38 - Laura Moyers

Planning meeting 24 August:

Attending: Aaron, Monica, Mark, Laura, Ranjeet

ARM has a relationship with USGS; Aaron developed the USGS-SDC DataONE MN, which was a custom implementation. Aaron is in the information gathering stage of a potential ARM MN development. We're looking at GMN (the Generic Member Node) as a potential MN software solution.

Aaron's experience as the SDC MN developer has given him experience interacting with DataONE, but a possible GMN solution for ARM offers a (hopefully) simpler interface with DataONE. If we chose GMN for ARM and it was successful, SDC might consider transitioning to GMN as well.

Monica is sharing her presentation about the benefits of becoming a MN and use of GMN as a solution. Here are the slides:

https://docs.google.com/presentation/d/1g5g919PruDKnOJWATcKtePsNj5j6821uPZ0hpyWwmmw/edit?usp=sharing

In general, repositories can choose to expose metadata-only, or metadata and data (content). The repository/MN can choose to replicate its data out to other DataONE MNs or not (there are reasons for and for not replicating depending on the repository's operating procedures).

ARM has it's own data discovery tool similar to a data checkout system; the datasets can be quite large, so the users are directed back to the ARM website to download content. See https://www.archive.arm.gov/discovery/ and https://www.archive.arm.gov/discovery/#v/results/s/.

For the potential ARM MN, metadata only would be exposed and the user directed back to the ARM data discovery page to access the content.

ARM metadata is in a huge database; this metadata can be exported into desired formats such as ISO, FGDC, etc.

65K measurements that ARM instruments collect data from

> 3000 data streams, updated daily

Metadata could be generated either at the data stream or measurement (or instrument?) level. This is a decision to be made by ARM.

Aaron inquired if we'd worked with RedHat in terms of MN implementation before. Yes we have, and Laura will connect Aaron with the other MN (or MNs) who've done RedHat.

ARM data is distributed to OSTI who then distributes it to data.gov. QUESTION: what metadata format is used to send to OSTI?

Next steps:

> Look at metadata, metadata generation, etc. (must be schema valid)
> Consider granularity of data (by instrument, by stream, by ????)

#15 - 2017-08-24 22:45 - Monica Ihli

Contact Persons:
* Ranjeet Devarakonda - devarakondar@ornl.gov; Primary Contact
* Stokes, Aaron M. - stokesam@ornl.gov

Notes about ARM Systems:
* Repository Domain: https://www.arm.gov/
* Science data files are very large. The preferred method of access is for end users who discover data of interest through DataONE to be linked to a landing page rather than a direct downloadable link, where they will retrieve the data through the ARM interface.
* ARM is using a "data checkout" system.
* Science metadata is stored in a metadata database.
* Science metadata files are programmatically generated rather than stored as static objects.
* Metadata format choice is flexible but includes ISO varieties and FGDC.
* Regarding metadata granularity: There are approximately

- 65,000 instruments
- 3,000 data streams
- 300 sites
- Each of these is an option for the granularity of a metadata document. At this time we are looking at the level of the data stream. We just need to make sure that the transition from the DataONE search result to the ARM system makes sense to users.

Next Steps:
* Generate a few sample metadata files.
* Determine the appropriate landing pages users would be redirected to for these sample metadata files.
* Evaluate if the metadata granularity and landing pages make sense from an information retrieval perspective, and from a user's perspective who must retrieve data based on the page to which they are sent from the dataone search interface.

#16 - 2017-08-29 08:29 - Laura Moyers

epad notes for future use/reference

https://epad.dataone.org/pad/p/ARM_and_DataONE

#17 - 2017-08-29 11:49 - Laura Moyers

Original "start" date 10/1/13. Technical conversations began at 8/17/17 MNF (general queries from Aaron at 7/20/17 MNF).

#18 - 2017-08-30 20:37 - Laura Moyers

Latest from Aaron 29 August 2017: We are currently in the process of discussing this opportunity with our sponsor, but their response will likely take some time. I'll make sure you're informed once we have word back from them.

Depending upon the outcome of discussions with sponsor, we will either proceed or defer this activity until some future time. Status=planning until we hear back from Aaron.

#19 - 2018-01-22 21:27 - Amy Forrester

epad NOTES CONSOLIDATION

11/6/17: Probably not to continue. Needs some outreach work, but may be dead in the water wrt ornl involvement
*Rebecca contact at AGU??

#20 - 2018-01-29 20:02 - Amy Forrester

- % Done changed from 10 to 100

- Status changed from Planning to Closed

#21 - 2018-05-10 20:37 - Amy Forrester

- % Done changed from 100 to 0

- Status changed from Closed to New

5/10/18 - Meeting @ ORNL -- Giri expressed interest to become MN in the Fall (see notes Story #8589)

#22 - 2018-05-14 19:22 - Amy Forrester

- % Done changed from 0 to 10

- Status changed from New to Planning

#23 - 2019-05-29 11:14 - Dave Vieglais

- Assignee changed from Monica Ihli to Roger Dahl

#24 - 2019-05-29 11:23 - Dave Vieglais

After discussion with Ranjeet and Giri, the goal is to access the repository using the schema.org approach. The metadata is stored in a database and generated on demand in the requested format.

Examples with ISO 19115-2 metadata:

Landing page with json-ld: https://www.archive.arm.gov/metadata/adc/html/nsadlfptC1.b1.html

Metadata: https://www.archive.arm.gov/metadata/adc/xml/nsadlfptC1.b1.xml

Landing page with json-ld: https://www.archive.arm.gov/metadata/adc/html/nsamfrsrC1.b1.html

Metadata: https://www.archive.arm.gov/metadata/adc/xml/nsamfrsrC1.b1.xml

The GMN instance will harvest from the sitemap advertised schema.org endpoints, generate system metadata, and present the ARM repository as a Member Node.

#25 - 2019-06-11 19:44 - Roger Dahl

- Assignee changed from Roger Dahl to John Evans

#26 - 2019-06-11 20:36 - John Evans

Progress so far:

I took the existing schema_org code as a starting point and fleshed out a harvester for IEDA, which seems to be able to parse and harvest all 546 documents found on the IEDA site map.

The differences with ARM are that ARM doesn't follow the sitemap.org conventions (their sitemap is just a flat file of URLS rather than an XML document with last modification times. The bigger problem is that the JSON-LD parsed out of the ARM HTML documents do not provide a link to the XML document.

The VLab entry from 2019-05-29 by Dave gives metadata URLs that seem to correspond to the URLs of the HTML documents in the site map, i.e. https://www.archive.arm.gov/metadata/adc/html/nsadlfptC1.b1.html is easily transformed into https://www.archive.arm.gov/metadata/adc/xml/nsadlfptC1.b1.xml, but that mapping doesn't seem to always hold. In fact, none of the HTML docs from the sitemap have the 'adc' string in them. Even accounting for that, only a handful HTML URLs transformed correctly into valid XML URLs.

This probably just needs discussion with Ranjeet and Giri.

#27 - 2019-06-17 13:40 - John Evans

Have started trying to harvest ARM records. The new sitemap contains 320 records.

Immediately ran into a problem where the DOI identifiers are not present in the JSON-LD. The '@id' identifiers now look like they only contain the string "http://dx.doi.org/". Right now this doesn't seem to be an absolute showstopper because there appears to be a head/meta[@name="citation_doi"] tag outside of the JSON-LD that does give us what we need. But this is not what we were expecting, correct?

After compensating for the unexpected location of the DOI identifier, I was able to harvest 199 out of 320 records. Currently looking at the other failures.

#28 - 2019-06-17 18:11 - John Evans

Dumb question about DOIs... As I understood it, a DOI has to uniquely resolve to a single document? All 596 IEDA documents I parsed had a unique DOI, but there are a lot of multiple-occurrence DOIs in the ARM documents. For example, both of the following documents have the same DOI - 10.5439/1027372

- https://www.archive.arm.gov/metadata/adc/html/nsaqcrad1longC2.c2.html
- https://www.archive.arm.gov/metadata/adc/html/nsaqcrad1longC1.s1.html

Having the same DOI causes the client code to skip the create stage in the 2nd document (since the DOI / SID already exists in the system) and then fail in the update stage since the DOI/SID in the 2nd doc is the same as the first.

That is correct, yes? That I am to use the DOI as the SID (native system identifier)?

#29 - 2019-06-17 18:13 - John Evans

And the ARM folks have updated their landing pages to include the full DOI in the JSON-LD...

#30 - 2019-06-17 18:22 - John Evans

Here's a raw description of the error I see when the update fails. The XML response that comes back is as follows:

<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="/mn/templates/home.xsl"?>

Operation is denied. level="write", pid="4b5431e3d092c17b1e4c6027bee78764", active_subjects="public (primary)"

#31 - 2019-06-18 13:52 - John Evans

When the access policy is changed from 'read' to 'write', this results in the following (taken from the log file):

2019-06-18 09:43:15,421 INFO Updated 100 records.
2019-06-18 09:43:15,421 INFO Skipped 0 records.
2019-06-18 09:43:15,422 INFO Created 199 new records.
2019-06-18 09:43:15,422 INFO Could not process 21 records.

I'm not 100% certain that changing the access policy was the right course of action, at least for the first document that was 'updated'. The two objects have the same DOI, 10:5439/1027372. Their XML docs are very similar, but differ in geographic extent. There's also a MD_ProgressCode in the documents, the first document has "completed" while the 2nd document has "onGogin" (presumably "ongoing"?). But it was the "ongoing" document that ends up overwriting the "completed" document, so some more smarts might be required here.

Looking into the 21 documents that could not be processed.

#32 - 2019-06-18 14:15 - John Evans

The 21 remaining documents that could not be processed were all cases of bad XML.

There are two different cases of bad XML, although both look like misuse of the ampersand. One case is pretty simple, there are a lot of bare ' & ' in the title element of each of the following documents, something like

```
Ten Meter Tower: meteorological data, 2 & 6 m, 1-min avg at Central Facility, Atqasuk AK - nsamettiptwrC2.a1
```

- https://www.archive.arm.gov/metadata/adc/xml/nsamfrsraod1michC2.s1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsanimfraod1michC1.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamwrlosC2.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsanimfraod1michC2.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamettiptwrC2.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamwrlosC1.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamwrlosC1.a1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsawsicloudC2.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamfrsraod1michC1.s1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamfrsraod1michC1.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsawsicloudC1.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsanimfraod1michC1.s1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamettwrC1.a1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsanimfraod1michC2.s1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamettwrC1.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamfrsraod1michC2.c1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsamettiptwrC2.a1.xml

The second case is similar. Four documents have a title string that look something like

```
Balloon-borne sounding system (BBSS): Vaisala-processed winds, press., temp, &RH at Central Facility, Barrow A
K - nsasondewnpnC1.a1
```

- https://www.archive.arm.gov/metadata/adc/xml/nsasondewnpnS01.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsasondewnpnC1.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsasondewnpnS02.b1.xml
- https://www.archive.arm.gov/metadata/adc/xml/nsasondewnpnC1.a1.xml

Again, improper use of the ampersand.

#33 - 2019-06-18 20:14 - John Evans

Have made a temporary update to check the contents of the MD_ProgressCode elements between the existing doc and the update doc. Only if the update doc shows 'completed' will the update be allowed to continue.

This is temporary and cannot run in production because it uses a raw requests session to get the existing document out of the D1 instance. This bit of code will have to be replaced.

The change allowed the full set of 320 records to be processed as follows:

```
2019-06-18 15:52:36,001 INFO     Updated 70 records.
2019-06-18 15:52:36,001 INFO     Skipped 0 records.
2019-06-18 15:52:36,001 INFO     Created 199 new records.
2019-06-18 15:52:36,001 INFO     Rejected 30 records.
2019-06-18 15:52:36,001 INFO     Could not process 21 records.
```

Presumably 30 records hit the case where the progress code was not sufficient to allow the update. Some additional spot checks should be performed to be sure this is working as expected.

#34 - 2019-06-19 14:44 - John Evans

Have confirmed that all 30 of the rejected record updates were caused by the existing document having an MD_Progress code of "complete" while the proposed updating document had a misspelled MD_Progress code of "onGogin" (presumed intent of "ongoing").

Of the 70 records that were successfully updated, 8 were cases of the existing document having a code of "onGogin" and the proposed updating document having a code of "completed". This seems right.

The case of the other 62 records is a bit less clear. In all such examples, both documents had codes of "completed". All cases had just minor differences in the XML, but one such difference was in the gmd:dateStamp element, which is the time of the last metadata update. We could judge

the later gmd:dateStamp as identifying which document we want.

#35 - 2019-06-21 15:24 - John Evans

So it turns out that ALL of the ARM metadata records are invalid. When validation is directly hooked into the harvest process by import XSD documents from https://data.noaa.gov/resources/iso19139/, every single ARM document fails to validate.

As an example, the document nsacmhC2.a1.xml (landing page https://www.archive.arm.gov/metadata/adc/html/nsacmhC1.a1.html) issues multiple errors.

```
$ xmllint --schema /Users/johnevans/tmp/try3/data.noaa.gov/resources/iso19139/schema.xsd tmp/nsacmhC1.a1.xml 1
> /dev/null
```

This raises three errors, although there are more that show up when one tries to address the errors).

```
1.  element CI_ResponsibleParty: Schemas validity error : Element '{http://www.isotc211.org/2005/gmd}CI_Respon
sibleParty', attribute 'id':
'person.8' is not a valid value of the atomic type 'xs:ID'.

This is apparently because the ids must be unique in the document, and this ID shows up twice.

2.  element CI_ResponsibleParty: Schemas validity error : Element
'{http://www.isotc211.org/2005/gmd}CI_ResponsibleParty': This element is not expected.

Apparently cannot have more than one of these in a row?  In this case, there are two
CI_ResponsibleParty elements.

3. element UnitDefinition: Schemas validity error : Element '{http://www.opengis.net/gml}UnitDefinition':
This element is not expected. Expected is one of (
 {http://www.opengis.net/gml/3.2}UnitDefinition,
{http://www.opengis.net/gml/3.2}BaseUnit,
{http://www.opengis.net/gml/3.2}DerivedUnit,
{http://www.opengis.net/gml/3.2}ConventionalUnit,
{http://www.isotc211.org/2005/gmx}ML_BaseUnit,
{http://www.isotc211.org/2005/gmx}ML_DerivedUnit,
{http://www.isotc211.org/2005/gmx}ML_ConventionalUnit,
{http://www.isotc211.org/2005/gmx}ML_UnitDefinition ).

The XML declaration must say "gml/3.2", not just "gml"
```

All of the IEDA records validated on my local mnode.

#36 - 2019-07-02 15:18 - John Evans

Had sent word of this back to Ranjeet Devarakonda way back on June 24, no word back yet. One of his associates, Giri Prakash, is out of the office until July 12.

#37 - 2019-11-19 19:10 - Dave Vieglais

Test instance NodeIdentifier = "urn:node:mnTestARM"

#38 - 2019-11-25 20:10 - Dave Vieglais

- Date Upcoming set to 2019-11-25

- MN Description set to ARM focuses on obtaining continuous measurements—supplemented by field campaigns—and providing data products that promote the advancement of climate models. Serving users worldwide, the ARM Data Center collects and archives approximately 25 terabytes of data per month. ARM data include routine data products, value-added products (VAPs), field campaign data, complementary external data products from collaborating programs, and data contributed by ARM principal investigators for use by the scientific community. Data quality reports, graphical displays of data availability/quality, and data plots are also available from the ARM Data Center.

- Name set to ARM - Atmospheric Radiation Measurement Research Facility

- Logo URL set to https://raw.githubusercontent.com/DataONEorg/member-node-info/master/production/graphics/web/ARM.png

- Information URL set to https://www.archive.arm.gov/