Member Nodes - Task #3855

MNDeployment # 3118 (Operational): Dryad Member Node

Dryad schema instance files fail to validate due to import resolution

2013-06-28 19:00 - Chris Jones

Status:	Closed	**Start date:**	2013-06-28
Priority:	Normal	**Due date:**	
Assignee:	Chris Jones	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	Operational		
Story Points:			

Description

When trying to validate a Dryad science metadata instance document (see #3854), the validation fails due to unresolvable schema locations for the dcterms and bibo namespaces. Likewise, the bibo namespace resolves to an OWL ontology rather than an XML schema file. We need to find a solution to the schema resolution issues so that the CNs can validate Dryad science metadata documents during synchronization.

A possible solution is to update the Dryad Schema (inducing a v3.2) that correctly provides schemaLocations for the namespaces declared. Since the bibo namespace points to an OWL Ontology, there's really no way to validate documents using XML Schema parsers. This would have to be resolved.

Another possible solution is to register the dcterms and bibo schema files with Metacat manually. However, there is no bibo.xsd file associated with the Bibliographic Ontology project at http://purl.org/ontology/bibo. I do see that Dryad has created a subset of the bibo ontology as an XML schema document (see "bibo.xsd": https://code.google.com/p/dryad/source/browse/trunk/dryad/dspace/modules/xmlui/src/main/webapp/themes/Dryad/meta/schema/v3/bibo.xsd?r=3567), but this does not represent the bibo namespace completely. The namespace should be changed to a dryad-subset of the Bibliographic Ontology, rather than representing the whole namespace with a subset schema.

Another possible solution is to modify the instance documents to explicitly include an xsi:schemaLocation attribute for each of the dcterms and bibo namespaces. Metacat would then pull the schema files from those locations and register them, although I'd need to confirm that in the code. It certainly resolves schema documents from schemaLocation declarations in XML Schema documents - I'm just not sure about the instance documents. This solution would still need to address the bibo subset namespace issue above.

History

#1 - 2013-06-28 21:09 - Chris Jones

Here's the link to the subset of the bibo namespace in the v3.1 branch:

https://raw.github.com/datadryad/dryad-repo/dryad-master/dspace/modules/xmlui/src/main/webapp/themes/Dryad/meta/schema/v3.1/bibo.xsd

I had seen the pointer to v3 in the code.google.com above, so I asume this is more recent.

#2 - 2013-06-28 21:46 - Ryan Scherle

- Status changed from New to In Progress

I corrected the files so the XML validates.

Regarding the status of the bibo schema... We originally decided to provide a partial XSD for bibo, because we're not attempting to be the "official" place to validate bibo. If there ever is an official bibo schema, we could change our dryad.xsd to point to it without affecting any validation. This seems like a non-issue to me, but maybe I'm just not understanding the problem.

#3 - 2013-07-01 17:18 - Chris Jones

Hi Ryan,

The issue with the bibo namespace is that "http://purl.org/ontology/bibo/" represents all of the concepts in the namespaceprovided by the Bibliographic Ontology project. If we register that namespace with your subset XSD, any future use (potentially not from Dryad, but other DataONE groups) that includes concepts in the namespace but not represented in your subset schema would fail to validate. So, although your subset conceptually aligns with a subset of bibo, we wouldn't be able to register your subset XSD as the authoritative schema for the namespace.

I think the best thing to do is to change your namespace declaration to be something like xmlns:dbibo="http://purl.org/dryad-bibo/schema/terms/v3.1" and then provide a schemaLocation to your subset XSD so parsers can resolve it. In this way, if the Bib O group ends up creating a full-fledged XSD, you could switch your content over to use it, as could other D1 participants.

#4 - 2013-07-04 04:50 - Rob Nahf

Trying to create one of the Dryad metadata objects on a metacat member node to stimulate the same response the CNs will give, it ran into parsing problems, with Metacat returning a "fatal processing error"

Below is the PI in dev.datadryad.org, the metadata itself, and the stack trace. It looks like something more is going on that we thought.

20130703-22:44:21: [WARN]: Original PID: http://dx.doi.org/10.5061/dryad.12?ver=2011-08-02T16:00:05.530-0400 [org.dataone.integration.it.MNCreateTest]

dcterms:typepackage/dcterms:type
dcterms:creatorZmasek, Christian M./dcterms:creator
dcterms:creatorZhang, Qing/dcterms:creator
dcterms:creatorYe, Yuzhen/dcterms:creator
dcterms:creatorGodzik, Adam/dcterms:creator
dcterms:dateSubmitted2007-12-06T20:55:28Z/dcterms:dateSubmitted
dcterms:available2007-12-06T20:55:28Z/dcterms:available
dcterms:titleData from: Surprising complexity of the ancestral apoptosis network/dcterms:title
dcterms:identifierhttp://dx.doi.org/10.5061/dryad.12/dcterms:identifier
dcterms:referenceshttp://dx.doi.org/10.1186/gb-2007-8-10-r226/dcterms:references
bibo:JournalGenome Biology/bibo:Journal
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/1/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/2/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/3/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/4/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/5/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/6/dcterms:hasPart

20130703-22:44:22: [WARN]: formatId: http://datadryad.org/profile/v3.1 [org.dataone.integration.it.MNCreateTest]
checkServerTrusted - RSA
org.dataone.service.exceptions.ServiceFailure: Error inserting or updating document: <?xml version="1.0"?>Fatal processing error.
at org.dataone.service.util.ExceptionHandler.deserializeXml(ExceptionHandler.java:624)
at org.dataone.service.util.ExceptionHandler.deserializeXmlAndThrowException(ExceptionHandler.java:508)
at org.dataone.service.util.ExceptionHandler.deserializeAndThrowException(ExceptionHandler.java:364)
at org.dataone.service.util.ExceptionHandler.deserializeAndThrowException(ExceptionHandler.java:314)
at org.dataone.service.util.ExceptionHandler.filterErrors(ExceptionHandler.java:108)
at org.dataone.service.util.ExceptionHandler.filterErrors(ExceptionHandler.java:83)
at org.dataone.client.D1RestClient.doPostRequest(D1RestClient.java:300)
at org.dataone.client.MNode.create(MNode.java:505)
at org.dataone.client.MNode.create(MNode.java:473)

with the logs (catalina.out) yielding:

knb 20130704-04:41:29: [DEBUG]: Allowed to insert: testDryadMD [edu.ucsb.nceas.metacat.dataone.D1NodeService]
knb 20130704-04:41:29: [DEBUG]: Starting to insert xml document... [edu.ucsb.nceas.metacat.dataone.D1NodeService]
knb 20130704-04:41:29: [FATAL]: DBSaxHandler.fatalError - White spaces are required between publicId and systemId. [edu.ucsb.nceas.metacat.DBSAXHandler]
knb 20130704-04:41:29: [ERROR]: DocumentImpl.write - Problem with parsing: Fatal processing error. [edu.ucsb.nceas.metacat.DocumentImpl]
org.xml.sax.SAXParseException: White spaces are required between publicId and systemId.
at edu.ucsb.nceas.metacat.DBSAXHandler.fatalError(DBSAXHandler.java:734)
at org.apache.xerces.util.ErrorHandlerWrapper.fatalError(Unknown Source)
....

#5 - 2013-07-04 16:32 - Ryan Scherle

- % Done changed from 0 to 100

After some discussion with the Dryad team, we decided to create a Dryad-specific namespace. This namespace is now in place for all science metadata on our dev system. We're ready for you to re-attempt a harvest.

#6 - 2013-07-05 15:30 - Rob Nahf

retrying duplicating under a new name on mnDemo5, I get essentially the same exception.
Here's the metadata document text:

20130705-09:05:41: [WARN]: Original PID: http://dx.doi.org/10.5061/dryad.12?ver=2011-08-02T16:00:05.530-0400 [org.dataone.integration.it.MNCreateTest]

dcterms:typepackage/dcterms:type
dcterms:creatorZmasek, Christian M./dcterms:creator
dcterms:creatorZhang, Qing/dcterms:creator
dcterms:creatorYe, Yuzhen/dcterms:creator
dcterms:creatorGodzik, Adam/dcterms:creator
dcterms:dateSubmitted2007-12-06T20:55:28Z/dcterms:dateSubmitted
dcterms:available2007-12-06T20:55:28Z/dcterms:available
dcterms:titleData from: Surprising complexity of the ancestral apoptosis network/dcterms:title
dcterms:identifierhttp://dx.doi.org/10.5061/dryad.12/dcterms:identifier
dcterms:referenceshttp://dx.doi.org/10.1186/gb-2007-8-10-r226/dcterms:references
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/1/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/2/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/3/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/4/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/5/dcterms:hasPart
dcterms:hasParthttp://dx.doi.org/10.5061/dryad.12/6/dcterms:hasPart

20130705-09:05:43: [WARN]: formatId: http://datadryad.org/profile/v3.1 [org.dataone.integration.it.MNCreateTest]
checkServerTrusted - RSA
org.dataone.service.exceptions.ServiceFailure: Error inserting or updating document: <?xml version="1.0"?>Fatal processing error.
at org.dataone.service.util.ExceptionHandler.deserializeXml(ExceptionHandler.java:624)

chris: well, there's two things that I see:
chris: 1) The 2 dryad schemas are resovable, but they go through 2 http redirects. Metacat may be stumbling on those, not sure.
chris: 2) The dcterms namespace is not resolvable at the namespace URI, which is fine, but there's no schemaLocation for it either, so I bet Metacat can't find it
chris: Now that they've changed the bibo schema location, there's no problem with us manually caching these schemas on the CNs
chris: not ideal, but it's fine
rob: chris, should we try to get them to specify a resolvable dcterms location?
chris: Given our time frame and all the back and forth, I'm inclined to manually cache them now, and then pursue that later. (I think we need to figure out what's required for new metadata schemas to be supported)

#7 - 2013-07-08 18:20 - Chris Jones

I've registered the following schemas in the Metacat instance on mn-demo-5.test.dataone.org in order to try a test create() which will validate the Dryad science metadata:

public_id | system_id

---+---
http://purl.org/dryad/schema/terms/v3.1 | /schema/dryad.xsd
http://purl.org/dc/terms/ | /schema/dcterms.xsd
http://purl.org/dc/elements/1.1/ | /schema/dc.xsd
http://purl.org/dc/dcmitype/ | /schema/dcmitype.xsd
http://purl.org/dryad/schema/dryad-bibo/v3.1 | /schema/bibo.xsd

On calling create(), I got the following exception:

org.dataone.service.exceptions.ServiceFailure: Error inserting or updating document: <?xml version="1.0"?>src-import.3.1: The namespace attribute, 'http://purl.org/ontology/bibo/', of an element information item must be identical to the targetNamespace attribute, 'http://purl.org/dryad/schema/dryad-bibo/v3.1', of the imported document.
at org.dataone.service.util.ExceptionHandler.deserializeXml(ExceptionHandler.java:624)
at org.dataone.service.util.ExceptionHandler.deserializeXmlAndThrowException(ExceptionHandler.java:508)
at org.dataone.service.util.ExceptionHandler.deserializeAndThrowException(ExceptionHandler.java:364)
at org.dataone.service.util.ExceptionHandler.deserializeAndThrowException(ExceptionHandler.java:314)
at org.dataone.service.util.ExceptionHandler.filterErrors(ExceptionHandler.java:108)
at org.dataone.service.util.ExceptionHandler.filterErrors(ExceptionHandler.java:83)
at org.dataone.client.D1RestClient.doPostRequest(D1RestClient.java:300)
at org.dataone.client.MNode.create(MNode.java:505)
at org.dataone.client.MNode.create(MNode.java:473)
at org.dataone.tests.MNCreateTest.main(MNCreateTest.java:50)

So, there's now a conflict between the bibo namespace declared in the instance documents, compared to the bibo namespace declared in the Dryad v3.1 schema. At this point, I think that the Dryad schema would need to be updated with the correct namespace, and then published as v3.2 (since v3.1 is already a published schema).

#8 - 2013-07-16 17:57 - Ryan Scherle

I don't see a problem here. The only system in the world that has "published" these 3.1 documents is the D1 dev system. Is it difficult to delete the old schema from the dev system and ingest the new one?

#9 - 2013-07-25 19:06 - Ryan Scherle

Ok, I found an error in our schema and corrected it. Please test this again.

#10 - 2013-07-25 19:18 - Ryan Scherle

- Assignee changed from Ryan Scherle to Chris Jones

(reassigning to Chris, since he's the most likely to test it.)

#11 - 2013-08-01 19:50 - Bruce Wilson

- Target version changed from Deploy by end of Y4Q4 to Deploy by end of Y5Q2

#12 - 2013-08-02 16:18 - Chris Jones

In testing the updated v3.1 schema, I'm getting the following error when trying to harvest Dryad science metadata content on cn-dev-ucsb-1:

[ERROR] 2013-08-02 16:12:26,036 (TransferObjectTask:write:457) Task-urn:node:mnTestDRYAD-
http://dx.doi.org/10.5061/dryad.18?ver=2011-08-02T17:02:49.844-0400
<?xml version="1.0" encoding="UTF-8"?>

Error inserting or updating document: <?xml version="1.0"?><error>cvc-complex-type.2.4.b: The content of element 'DryadDataPackage' is not complete. One of '{"http://purl.org/dc/terms/":relation, "http://purl.org/dc/terms/":references}' is expected.</error>

I'm tracking this down now. This is encouraging in that we may have resolved the schema issues, and this may just be a content issue in the instance files (hopefully).

#13 - 2013-08-12 18:55 - Ben Leinfelder

- Assignee changed from Chris Jones to Ryan Scherle

Hi Ryan,
We are still having trouble validating your schema. The last error message:

cos-nonambig: "http://purl.org/dc/terms/":relation and "http://purl.org/dc/terms/":references (or elements from their substitution group) violate "Unique Particle Attribution". During validation against this schema, ambiguity would be created for
those two particles.

indicates that there is the possibility for ambiguity because the "references" element is a substitute for "relation" -- defining them both side-by-side means a parser cannot determine if the instance doc has a true "relation" element (optional) or a "relation" element as a substitute for "references" (required).

I've tried modifying dryad.xsd to remedy this, and it is possible. Basically you have to ensure that no "relation" or relation-substitutes are next to each other (either directly or because they are separated by option (minOccurs="0") elements.

Here's the definition I ended up with for "DryadDataPackage":

<!-- For the Dryad data package -->

xs:complexType
xs:sequence

/xs:sequence
/xs:complexType
/xs:element

#14 - 2013-08-12 21:31 - Ryan Scherle

- Assignee changed from Ryan Scherle to Chris Jones

After some analysis, I determined that we are no longer using the ambiguous field "relation". I have removed it from the schema.

The documents validate in oXygen. However, they validated before this change. If you're still having validation problems, let me know...

#15 - 2013-08-13 19:29 - Skye Roseboom

- Assignee changed from Chris Jones to Ryan Scherle

Test run for sync'ing Dryad content - 8-13-13:

Majority of DryadDataPackages and DryadDataFiles synchronized - total 5001 out of 5496 science metadata files synchronized. So 495 documents failed to synchronize.

Encountered several classes of validation errors:
DryadDataPackage:

Schema requires one and only one dcterms:references element. Discovered instance documents that do not contain this element - for example (not an exhaustive list):
http://dx.doi.org/10.5061/dryad.150?ver=2011-08-30T15:07:52.418-0400
http://dx.doi.org/10.5061/dryad.152?ver=2012-02-27T11:00:36.289-0500
http://dx.doi.org/10.5061/dryad.8r7s5?ver=2012-05-18T16:02:32.255-0400

DryadDataFiles:

Schema requires one and only one 'dcterms:rights'. Discovered instance documents that do not have any dctemrs:rights elements and some that have more than one dc:terms rights elements - for example (not an exhaustive list):
http://dx.doi.org/10.5061/dryad.1916/1?ver=2010-08-27T15:15:36.655-0400 (missing)
http://dx.doi.org/10.5061/dryad.1914/1?ver=2010-10-12T16:16:13.676-0400 (missing)
http://dx.doi.org/10.5061/dryad.9063/1?ver=2011-08-02T11:21:47.834-0400 (too many)
http://dx.doi.org/10.5061/dryad.b779h/1?ver=2011-12-20T15:00:30.966-0500 (too many)

Schema requires at least one 'dcterms:creator' element. Discovered instance documents that do not have this element. For example:
http://dx.doi.org/10.5061/dryad.1998/2?ver=2010-12-06T16:43:48.805-0500
http://dx.doi.org/10.5061/dryad.1968/1?ver=2010-11-16T11:22:07.577-0500
http://dx.doi.org/10.5061/dryad.7908/1?ver=2010-12-23T15:09:25.968-0500

#16 - 2013-09-24 20:52 - Ryan Scherle

- Assignee changed from Ryan Scherle to Chris Jones

These issues are now fixed on the Dryad dev server.

#17 - 2014-02-03 16:03 - Laura Moyers

- Target version changed from Deploy by end of Y5Q2 to Deploy by end of Y5Q3

#18 - 2014-03-14 18:22 - Laura Moyers

- Target version changed from Deploy by end of Y5Q3 to Operational

#19 - 2014-04-17 03:31 - Chris Jones

- translation missing: en.field_remaining_hours set to 0.0

- Status changed from In Progress to Closed