Member Nodes - Task #5940

MNDeployment # 3708 (Operational): Minnesota Population Center

Task # 5921 (Closed): MPC: Testing

Task # 5922 (Closed): MPC: Registration in environment

Task # 5933 (Closed): MPC: Content Review

Task # 5937 (Closed): MPC: Verify Science Metadata

MPC: Verify that the Science Metadata is correctly processed by CNs

2014-07-18 15:31 - Laura Moyers

Status:	Closed	**Start date:**	2014-07-18
Priority:	Normal	**Due date:**	
Assignee:	Chris Jones	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	Operational		
Story Points:			

Description

Verify that CNs are able to index any supported Science Metadata objects.
Successfully indexed objects appear in the Solr index. Not all Science Metadata formats are supported by DataONE. If the MN does not expose any supported objects, this ticket can be closed directly.

Related issues:

Related to Infrastructure - Bug #6477: Verify MPC QDC Science Metadata files ...	**Rejected**

History

#1 - 2014-08-04 18:47 - Laura Moyers

- Target version changed from Deploy by end of Y5Q4 to Deploy by end of Y1Q1

#2 - 2014-09-19 13:51 - Chris Jones

- Status changed from New to In Progress

- Assignee set to Chris Jones

I'll be turning on synchronization in the stage environment to verify content processing. Stand by.

#3 - 2014-09-23 18:58 - Chris Jones

In harvesting science metadata content to the stage Coordinating Node, we've run into parsing errors for the Dublin Core metadata files:

<?xml version="1.0"?>

cvc-elt.1: Cannot find the declaration of element 'qualifieddc'.

In looking at an example XML file like https://dataone-test.pop.umn.edu/mn/v1/object/ipumsi_6.3_ke_1999_DC.xml, I see two possible issues causing the parsing problem:

1) The element insn't namespaced. While the document's schema location is provided, there's no xmlns declaration for the qualifieddc schema, and I'm thinking the parser doesn't know that is an element defined in qualifieddc.xsd. I think all of the documents need to be updated to declare the namespace of the root element, like

...
/qdc:qualifieddc

In this way, the root element should be parsed correctly.

2) The xsi:schemaLocation attribute has:

xsi:schemaLocation="qualifieddc.xsd http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd"

I think the values in this field are reversed. It should be a pair of {NAMESPACE, LOCATION} strings, with the namespace first, like:

xsi:schemaLocation="http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd qualifieddc.xsd"

So, the parser should know that this document is adhering to the 'http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd' namespace, and it will find the physical schema at the location given, or in it's catalog. For the CNs, it will find it in the schema catalog.

So, I think that addressing these two issues should solve the parsing issue found above. I'll send this to Wendy and Fabio to have a look.

#4 - 2014-09-24 01:58 - Chris Jones

Wendy and Fabio updated their science metadata documents with the above changes. We unfortunately now have a new error:

Error inserting or updating document: <?xml version="1.0"?><error>TargetNamespace.1:
Expecting namespace 'http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd',
but the target namespace of the schema document is 'null'.</error>

My thought here is that the schema validators on my workstation (xmllint, xmlstarlet, both using libxml) are configured differently than the Xerces SAX parser on the CNs. Documents that had un-namespaced elements validated fine on my machine, whereas the Xerces SAX parser throws the error above. When I change the document locally to use qdc:qualifieddc, xmllint throws an error, saying:

ipumsi_6.3_am_2001_DC.xml:2: element qualifieddc:
Schemas validity error : Element '{http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd}qualifieddc':
No matching global declaration available for the validation root.
ipumsi_6.3_am_2001_DC.xml fails to validate

This points to the fact that the qualifieddc schema downloaded from http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd doesn't have a targetNamespace attribute, since, according to DCMI, this is an un-namespaced 'container' schema. This behavior matches what we see from Xerces, which says that the target namespace for this schema is 'null'.

So, we're stuck between a rock and a hard place. Xerces doesn't like un-namespaced elements like , but when you namespace it with qdc:qualifieddc, it somewhat flippantly tells us that the schema used to validate it isn't typed with that namespace. Ugh, software!

Perhaps there is a property setting on the Xerces parser that allows untyped root elements, but typed child elements? Am I missing something else? I'll ping others (Ben, Matt, Dave especially) to get more eyes on this.

#5 - 2014-09-29 21:30 - Laura Moyers

- Target version changed from Deploy by end of Y1Q1 to Deploy by end of NCTE

#6 - 2015-01-02 16:42 - Laura Moyers

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

#7 - 2015-01-28 15:33 - Laura Moyers

- % Done changed from 0 to 100

- translation missing: en.field_remaining_hours set to 0

- Status changed from In Progress to Closed