Metacat - Bug #7091

Metacat installation (the war file) doesn't include the FGDC schemas even though the loaddtdschema file creates an entry in the xml_catalog table for it

08/24/2016 04:12 PM - Jing Tao

Status:	New	**Start date:**	08/24/2016
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.9.0	**Spent time:**	0.00 hour
Bugzilla-Id:			

Description

The FGDC namespace is registered in our xml_catalog table:
27 | Schema | | | metadata | /schema/fgdc-std-001/fgdc-std-001-1998.xsd

`However, the schema location schema/fgdc-std-001/fgdc-std-001-1998.xsd doesn't exist.`

History

#1 - 08/24/2016 04:59 PM - Jing Tao

To register the entry on loaddtdschema-postgres.sql exists for a long time since we can see it on Metacat 1.9.0. However, we can't find the actually file on svn. The lib/schema begins on Metacat 2.4.0 and it doesn't have the directory.
I also downloaded metacat-bin-1,9.0, 2.0.1 and 2.4.0 zip file and couldn't find the directory ever exists. So I double it ever exists at all.
I talked with Ben and he couldn't find it either.

#2 - 08/24/2016 05:21 PM - Matt Jones

The lib/schema directory was historically populated by first use, rather than during the build. It was only later that we fixed the particular schema versions in the build. Nevertheless, some FGDC DTD and XSD documents of interest may be here:

https://github.com/NCEAS/eml/tree/master/other

Or maybe we pulled it from another location before during the build process? In any case, the official FGDC schema file for that version is here:

https://www.fgdc.gov/schemas/metadata/fgdc-std-001-1998.xsd

#3 - 08/25/2016 11:49 AM - Dave Vieglais

There are two common locations used for FGDC schemas, both appear to be functionally equivalent and yield identical schemas:

https://www.fgdc.gov/schemas/metadata/fgdc-std-001-1998.xsd

and

https://www.fgdc.gov/metadata/fgdc-std-001-1998.xsd

When processing FGDC documents, either location may be specified in xsi: noNamespaceSchemaLocation and hence mapping of either location to the local copy of schema files must be supported.

Note that this issue is affecting production systems in DataONE.

#4 - 08/25/2016 12:03 PM - Jing Tao

The problem is that even though we have a cached copy, we still need to modify the noNamespaceSchemaLocation attribute to point to the cached version. Otherwise it would not work. See https://redmine.dataone.org/issues/7870

The easiest solution for now is to remove the attributes xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" and xsi:noNamespaceSchemaLocation="http://www.fgdc.gov/metadata/fgdc-std-001-1998.xsd". It will work. Actually SEAD has lots of documents don't have the two attributes :)

#5 - 08/30/2016 04:34 PM - Jing Tao

Let's go back to the original issue. We have a record on xml_catalog table:

```
catalog_id | entry_type | source_doctype | target_doctype | public_id |              system_id               | format_id
------------+------------+----------------+----------------+-----------+--------------------------------------+-----------
27 | Schema     |                |                | metadata  | /schema/fgdc-std-001/fgdc-std-001-1998.xsd |
```

However, we can't find any fgdc schemas with the target namespace "metadata".

Matt provided a schema:
https://github.com/NCEAS/eml/blob/bc894b1a4feff18e264fd851a4753be169ac444c/other/nbii-fgdc-std-001.1-1999.xsd
However, tt has a target namespace "nbii-fgdc-std-001.1-1999.dtd" rather than "metadata".

I just checked if anyone ever used the namespace "metadata".

In knb, there is no record for the entry in xml_catalog table at all:
knb=> select * from xml_catalog where public_id='metadata';
catalog_id | entry_type | source_doctype | target_doctype | public_id | system_id | format_id
------------+------------+----------------+----------------+-----------+-----------+-----------
(0 rows)

in cn-ucsb-1:
metacat=> select * from xml_catalog where public_id='metadata';
catalog_id | entry_type | source_doctype | target_doctype | public_id | system_id | format_id
------------+------------+----------------+----------------+-----------+---+-----------
27 | Schema | | | metadata | /schema/fgdc-std-001/fgdc-std-001-1998.xsd |
metacat=> select * from xml_documents where catalog_id =27;
docid | rootnodeid | docname | doctype | user_owner | user_updated | server_location | rev | date_created | date_updated | public_access | catalog_id
-------+------------+---------+---------+------------+--------------+-----------------+-----+--------------+--------------+---------------+------------
(0 rows)

The cn-ucsb-1 has the schema, but no objects associate with it.

So I am thinking either we can leave as it is or change the namespace to "nbii-fgdc-std-001.1-1999.dtd" and use the schema which matt provided.

#6 - 08/30/2016 05:54 PM - Matt Jones

Jing -- I think at one point (and maybe currently) there was a chunk of code that used the name of the root element of an XML document as the DOCTYPE and DOCNAME, back in the DTD days. My memory is shaky on this, but I think the algorithm first checked if the doc had a namespace (and if so used the schema associated with that), then checked if the document has a DOCTYPE declaration (and used the schema/dtd), and then checked if there was a dtd/schema registered with the name of the root elelment of the document (which, in the case of FGDC, was "metadata"). It was common to get FGDC documents with no DOCTYPE and no namespace, and we used the root element to trigger validation. We can certainly revisit that policy decision, but while you're at it you should carefully check all of the conditional logic to see how Metacat falls through the various options for determining the type of a document and locating the schema. It would be nice to see a little pseudocode explanation of 1) how the logic works now/previously, and 2) how you'd like to see it work now. Something like:

If (namespaceDefinedOnRoot)
... do some stuff, like check if schema is registered for that namespace
else if DOCTYPE defined in PROLOG
... do some other stuff
else if (root element name is registered in xml_catalog)
... yet more stuff
else
... fall through of last resort

The logic should define exactly how we will determine the type of the document, how we will check if an appropriate schema or DTD exists, and how to ignore or handle schemaLocation and noNamespaceSchemaLocation and related attributes. Then we can review that approach to be sure it is right for our different use cases before you implement it.

#7 - 08/30/2016 06:31 PM - Jing Tao

- Subject changed from Couldn't find the cached schema file for FGDC to Metacat installation (the war file) doesn't include the FGDC schemas even though the loaddtdschema file creates an entry in the xml_catalog table for it

I changed the title of this bug. It is different to the case we tried to figure out the xml schema location. The correct ticket for that issue is:
https://projects.ecoinformatics.org/ecoinfo/issues/7098

Matt I will copy you last comment to ticket 7098.

#8 - 09/12/2016 03:44 PM - Jing Tao

- Target version changed from 2.7.2 to 2.8.0

#9 - 10/13/2016 02:53 PM - Jing Tao

- Target version changed from 2.8.0 to 2.9.0