

Infrastructure - Task #3867

ORE parsing error: ore:describes element

2013-07-18 22:36 - Skye Roseboom

Status:	Closed	Start date:	2013-07-18
Priority:	Normal	Due date:	
Assignee:	Rob Nahf	% Done:	100%
Category:	d1_libclient_java	Estimated time:	0.00 hour
Target version:		Story Points:	
Milestone:	None		
Product Version:	*		

Description

Testing foresite ORE document parsing on production member nodes discovered that ORE documents from Merritt and ONEShare MN raise errors in the foresite RDF/ORE parsing library.

The error is related to the [ore:describes](#) element as defined in ORE docs from these MN -- for example: <https://cn.dataone.org/cn/v1/resolve/ark%3A%2F13030%2Fm50000sp%2F1%2Fmrt-dataone-map.rdf>

Line 36:

[ore:describeshttp://store.cdlib.org:35121/content/1001/ark%3A%2F13030%2Fm50000sp/1//ore:describes](#)

Seems to be missing the rdf:resource definition - for example:

Once this modification is made, foresite parsing seems happy.

Need to determine if this issue can be resolved in parsing or whether these documents are actually valid RDF/ORE.

Related issues:

Related to Member Nodes - Task #3906: Update malformed Resource Maps

New

2013-08-09

History

#1 - 2013-07-18 22:39 - Skye Roseboom

- Description updated

#2 - 2013-07-23 20:53 - Rob Nahf

- Status changed from New to In Progress

- % Done changed from 0 to 30

The python foresite library doesn't have the same issue as the java foresite library. It treats object-literals that are valid URIs the same as it would things represented as rdf:resources. It seems to inherit this behavior from the python rdflib module, so it's not a conscious decision on its part.

The java version assumes it's getting an OREResource, and throws an exception when it tries to cast the literal as one. It seems to be overly strict.

The ORE user guide recommends using the rdf:resource syntax for URIs, but it's not a requirement. (see ObjectRule under <http://www.openarchives.org/ore/1.0/rdfxml#SummaryRDFXML>)

"If the object of the triple is a URI reference, add an attribute with the QName rdf:resource and make the value of this attribute a URI reference corresponding to the object"

#3 - 2013-07-24 21:42 - Rob Nahf

Dave pointed out that the reference RDF checker shows two disconnected sub-graphs when fed a Merritt resource map, which is contrary to ORE specifications. Since the main problem is with the indexer, updating the Merritt and ONEShare resource maps would obsolete and archive the existing ones that are causing problems, and bypass the problem.

Checking with Mark Reyes at Merritt to see if they are able to accomplish this.

Otherwise, we would need to resort to extending the Jena foresite implementation in `d1_libclient_java`, and override the behavior in that circumstance.

#4 - 2013-08-09 17:37 - Rob Nahf

- *Status changed from In Progress to Closed*

- *% Done changed from 30 to 100*

- *translation missing: en.field_remaining_hours set to 0.0*

these disconnected graphs should be considered as malformed resource maps. Merritt will create updates for their (and ONEShare's) bad maps.