Infrastructure - Story #7668

Determine how indexing of data packages should work

2016-03-02 00:16 - Bryce Mecum

Status:	New	**Start date:**	2016-03-01
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	CCI-2.4.0		
Story Points:			

Description

I've discovered (with Lauren's help) a strange requirement for how the resource maps for nested data packages have to be written. In order to get nested data packages correctly indexed in Solr so that the 'resourceMap' field of the resource map being nested is set to the parent resource map's PID, you have to create the appropriate set of @cito:documents@ statements in addition to the expected @ore:aggregates@ statements.

I expected the following to be sufficient (pardon the highly abstracted RDF, examples are linked below):

parent_resource_map#aggregation ore:aggregates child_resource_map
parent_resource_map#aggregation ore:aggregates metadata_object

but I also had to add a @cito:documents@ statement between the *parent resource map's metadata object* and the resource maps being nested

parent_resource_map#aggregation ore:aggregates child_resource_map
parent_resource_map#aggregation ore:aggregates metadata_object

parent_metadata_object cito:documents child_resource_map

The documentation does not suggest this and I found it confusing. A real life example of what I expected to work is here:
https://gist.github.com/amoeba/c7a6ba269c5a1f78db1d
What I actually had to insert is here:
https://dev.nceas.ucsb.edu/knb/d1/mn/v2/object/resourceMap_urn:uuid:ab17b047-a341-4d06-b433-92eed90dacec

Is the need for the @cito:documents@ statement(s) really required and is this the intended behavior? I've made this issue in the hopes we can talk about it.

I suggest updating the API docs with whatever we decide, and hopefully that update will include example RDF for a nested data package.

Related issues:

Related to Infrastructure - Task #3156: Design Review: resource map indexing ...	**New**	**2012-08-27**

History

#1 - 2016-11-04 00:13 - Jing Tao

- Category set to d1_cn_index_processor

- Assignee set to Jing Tao

- Target version set to CCI-2.3.1

#2 - 2016-11-04 00:25 - Chris Jones

My understanding is that there shouldn't be a requirement to add a

parent_metadata_object cito:documents child_resource_map

statement. To me, this documentation isn't correct:

??A data package in DataONE is composed of at least one science metadata document describing at least one data object with the relationships between them documented in a resource map document.??

See https://releases.dataone.org/online/api-documentation-v2.0/design/DataPackage.html#synopsis

The six requirements of a DataONE Data Package that are over and above the OAI-ORE spec are found here:

https://releases.dataone.org/online/api-documentation-v2.0/design/DataPackage.html#generating-resource-maps

None of these require a @cito:documents@ statement, and so I think this documentation needs to be updated. Likewise, the @resourceMapSubprocessor@ code needs to be reviewed to remove any hard dependency on the presence of a @cito:documents@ statement. Currently, the @resourceMap@ field in Solr represents the @ore:aggregates@ statements, and I think that that only should be used to indicate participation in a resource map. In fact, I'd even prefer having the @aggregates@ and @isAggregatedBy@ Solr fields instead to show both directions of the relationship in the index.

An example is where a data manager calls @MN.create()@ on a bunch of data objects, and then calls @MN.create()@ on the resource map with these objects aggregated. Later, when the data manager is able to get the metadata from a scientist, they may call @MN.create()@ for the science metadata document, and @MN.update()@ on the resource map, adding the science metadata to the aggregation, and inserting @cito:documents@ statements.

#3 - 2016-12-01 23:55 - Dave Vieglais

- Target version changed from CCI-2.3.1 to CCI-2.4.0

#4 - 2017-03-28 16:14 - Dave Vieglais

- Category changed from d1_cn_index_processor to d1_indexer

- Project changed from CN Index to Infrastructure

- Milestone set to None

#5 - 2017-04-26 20:32 - Rob Nahf

- Related to Task #3156: Design Review: resource map indexing strategy added

#6 - 2017-04-26 20:46 - Rob Nahf

- Tracker changed from Task to Story

#7 - 2018-01-17 19:28 - Dave Vieglais

- Sprint set to Infrastructure backlog