Infrastructure - Task #6843

Update the prov instance of the RdfXmlSubprocessor to index renamed and inverse provenance properties

2015-02-06 23:18 - Chris Jones

Status:	In Progress	**Start date:**	2015-02-06
Priority:	Normal	**Due date:**	
Assignee:	Chris Jones	**% Done:**	30%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	CCI-2.4.0		
Milestone:	None	**Story Points:**	
Product Version:			

Description

In the "sem-prov-design issue 66":https://github.com/DataONEorg/sem-prov-design/issues/66 we have renamed the provenance-based Solr fields to include 'prov_' as a prefix, and have added new fields. See also "issue 99": https://github.com/DataONEorg/sem-prov-design/issues/99 and "issue 100": https://github.com/DataONEorg/sem-prov-design/issues/100.
Modify the provRdfXmlSubprocessor bean to handle the renaming scheme, the new fields, and the inverse fields determined to be useful. Also, add these fields as static Solr fields so we can remove the '_sm' suffixes from the names.

Associated revisions

Revision 15183 - 2015-02-07 01:52 - Chris Jones

Change indexing of provenance-related fields to use field names with a 'prov_' prefix, and drop the dynamic field suffixes. Also change hadExecution to two fields: prov_wasExecutedByExecution and prov_wasExecutedByUser. refs #6843

Revision 15183 - 2015-02-07 01:52 - Chris Jones

Change indexing of provenance-related fields to use field names with a 'prov_' prefix, and drop the dynamic field suffixes. Also change hadExecution to two fields: prov_wasExecutedByExecution and prov_wasExecutedByUser. refs #6843

Revision 15186 - 2015-02-07 06:57 - Chris Jones

Rename hadExecution bean to prov.wasExecutedByExecution and prov.wasExecutedByUser. refs #6843

Revision 15186 - 2015-02-07 06:57 - Chris Jones

Rename hadExecution bean to prov.wasExecutedByExecution and prov.wasExecutedByUser. refs #6843

Revision 15190 - 2015-02-08 22:53 - Chris Jones

Modify the provRdfXmlSubprocessor bean to index the prov_hasDerivations field, which indexes science metadata documents that describe data entities that were sources of other data entities. Also fix the SPARQL query logic in the prov_hasSources bean. refs #6843

Revision 15190 - 2015-02-08 22:53 - Chris Jones

Modify the provRdfXmlSubprocessor bean to index the prov_hasDerivations field, which indexes science metadata documents that describe data entities that were sources of other data entities. Also fix the SPARQL query logic in the prov_hasSources bean. refs #6843

Revision 15209 - 2015-02-12 01:08 - Chris Jones

copy the provenance application context beans from d1_cn_index_processor to update the most recent field name changes in the SPARQL queries that get deployed. refs #6843

Revision 15209 - 2015-02-12 01:08 - Chris Jones

copy the provenance application context beans from d1_cn_index_processor to update the most recent field name changes in the SPARQL queries that get deployed. refs #6843

Revision 15210 - 2015-02-12 01:21 - Chris Jones

Pilot error - I missed adding this file to the last commit. refs #6843

Revision 15210 - 2015-02-12 01:21 - Chris Jones

Pilot error - I missed adding this file to the last commit. refs #6843

Revision 15212 - 2015-02-12 01:38 - Chris Jones

Update the fixed SPARQL query in the buildout. refs #6843

Revision 15212 - 2015-02-12 01:38 - Chris Jones

Update the fixed SPARQL query in the buildout. refs #6843

Revision 15213 - 2015-02-12 01:55 - Chris Jones

Fix attribute name typo: multiValued, not multivalued. refs #6843

Revision 15213 - 2015-02-12 01:55 - Chris Jones

Fix attribute name typo: multiValued, not multivalued. refs #6843

Revision 15214 - 2015-02-12 01:56 - Chris Jones

Fix attribute name typo: multiValued, not multivalued. refs #6843

Revision 15214 - 2015-02-12 01:56 - Chris Jones

Fix attribute name typo: multiValued, not multivalued. refs #6843

Revision 15219 - 2015-02-12 17:17 - Chris Jones

Uncomment the provRdfXmlSubprocessor. refs #6843

Revision 15219 - 2015-02-12 17:17 - Chris Jones

Uncomment the provRdfXmlSubprocessor. refs #6843

Revision 15235 - 2015-02-17 16:56 - Chris Jones

Add minor debugging to help track down identifiers referenced in resource maps not found in the Solr index. refs #6843

Revision 15235 - 2015-02-17 16:56 - Chris Jones

Add minor debugging to help track down identifiers referenced in resource maps not found in the Solr index. refs #6843

Revision 15237 - 2015-02-17 19:35 - Chris Jones

Add the application-context-annotator bean definition into the tests to get the index processor tests working on jenkins. refs #6843

Revision 15237 - 2015-02-17 19:35 - Chris Jones

Add the application-context-annotator bean definition into the tests to get the index processor tests working on jenkins. refs #6843

Revision 15250 - 2015-02-23 19:15 - Chris Jones

Modify the RdfXmlSubprocessor to use mergeWithIndexedDocuments() rather than mergeDocs() to do the merging of content with content already in the index. The SolrIndexService iterates through the results provided by processDocument(), and so will index all of the document identifiers extracted out of the triple statements found in the RDF/XML document. I've removed mergeDocs() (not needed now) and fleshed out mergeWithIndexedDocument(). refs #6843

Revision 15250 - 2015-02-23 19:15 - Chris Jones

Modify the RdfXmlSubprocessor to use mergeWithIndexedDocuments() rather than mergeDocs() to do the merging of content with content already in the index. The SolrIndexService iterates through the results provided by processDocument(), and so will index all of the document identifiers extracted out of the triple statements found in the RDF/XML document. I've removed mergeDocs() (not needed now) and fleshed out mergeWithIndexedDocument(). refs #6843

Revision 15372 - 2015-03-25 00:14 - Chris Jones

Minor changes to indexing prov_used and prov_wasGeneratedBy. refs #6843

Revision 15372 - 2015-03-25 00:14 - Chris Jones

Minor changes to indexing prov_used and prov_wasGeneratedBy. refs #6843

Revision 15385 - 2015-03-26 22:18 - Chris Jones

After troubleshooting issues that Lauren pointed out:

- Remove the prov_wasGeneratedBy Solr field (in favor of just prov_wasGeneratedByExecution and prov_wasGeneratedByProgram)
- Add the prov_generated field as the inverse, where a program generates an entity
- Update the testProvenanceFields() test to reflect the above
- For now, ignore the testInsertProvResourceMap() test due to HZ conflicts

refs #6843

Revision 15385 - 2015-03-26 22:18 - Chris Jones

After troubleshooting issues that Lauren pointed out:

- Remove the prov_wasGeneratedBy Solr field (in favor of just prov_wasGeneratedByExecution and prov_wasGeneratedByProgram)
- Add the prov_generated field as the inverse, where a program generates an entity
- Update the testProvenanceFields() test to reflect the above
- For now, ignore the testInsertProvResourceMap() test due to HZ conflicts

refs #6843

Revision 15386 - 2015-03-26 22:18 - Chris Jones

- Remove the prov_wasGeneratedBy Solr field (in favor of just prov_wasGeneratedByExecution and prov_wasGeneratedByProgram)
- Add the prov_generated field as the inverse, where a program generates an entity

refs #6843

Revision 15386 - 2015-03-26 22:18 - Chris Jones

- Remove the prov_wasGeneratedBy Solr field (in favor of just prov_wasGeneratedByExecution and prov_wasGeneratedByProgram)
- Add the prov_generated field as the inverse, where a program generates an entity

refs #6843

Revision 15387 - 2015-03-29 01:11 - Chris Jones

We've found a bug during index processing that causes documents to not be indexed because of a mismatch between a 'beginDate' or 'endDate' field being added, and an existing 'beginDate' or 'endDate' field. This is a temporary fix that converts potential date strings to Date objects, and compares them. If they match, it is a dupe, and we don't add the field. Otherwise we add it. See and refs #6843

Revision 15387 - 2015-03-29 01:11 - Chris Jones

We've found a bug during index processing that causes documents to not be indexed because of a mismatch between a 'beginDate' or 'endDate' field being added, and an existing 'beginDate' or 'endDate' field. This is a temporary fix that converts potential date strings to Date objects, and compares them. If they match, it is a dupe, and we don't add the field. Otherwise we add it. See and refs #6843

Revision 15388 - 2015-03-29 20:23 - Chris Jones

Update the v1_1 Solr schema file with the prov_ field changes.

refs #6843

Revision 15388 - 2015-03-29 20:23 - Chris Jones

Update the v1_1 Solr schema file with the prov_ field changes.

refs #6843

Revision 15389 - 2015-03-29 20:50 - Chris Jones

Update the other various versions of the Solr schema to reflect the prov_ changes, and add prov_generated and prov_used to the query field descriptions file. Again, add all prov_* to the list of default return fields forSolr queries.

refs #6843

Revision 15389 - 2015-03-29 20:50 - Chris Jones

Update the other various versions of the Solr schema to reflect the prov_ changes, and add prov_generated and prov_used to the query field descriptions file. Again, add all prov_* to the list of default return fields forSolr queries.

refs #6843

Revision 15456 - 2015-04-19 20:05 - Chris Jones

Fix merging issues in the RdfxmlSubprocessor where, if mergeDocs() is called twice, and the existing map is larger than the pending map, we don't drop all of the existing documents, but rather add them to the merged map. Also, implement mergeWithIndexeddocument() in the same way as the AnnotatorSubprocessor (note that by following the API, we are cumulatively merging all existing docs, which seems a bit inefficient). refs #6843

Revision 15456 - 2015-04-19 20:05 - Chris Jones

Fix merging issues in the RdfxmlSubprocessor where, if mergeDocs() is called twice, and the existing map is larger than the pending map, we don't drop all of the existing documents, but rather add them to the merged map. Also, implement mergeWithIndexeddocument() in the same way as the AnnotatorSubprocessor (note that by following the API, we are cumulatively merging all existing docs, which seems a bit inefficient). refs #6843

History

#1 - 2015-02-07 01:52 - Chris Jones

- Status changed from New to In Progress

- % Done changed from 0 to 30

#2 - 2015-02-11 04:43 - Chris Jones

This is finished, and we are now indexing fields with prov_ prefixes.

#3 - 2015-03-16 17:54 - Chris Jones

Lauren pointed out that the prov_used and prov_wasGeneratedBy Solr fields should be populated by the Program that used or generated the Entity, rather than the Execution, since Executions aren't first-class objects per se in DataONE. I'm changing the SPARQL queries for these fields to reflect this.

#4 - 2015-03-29 01:24 - Chris Jones

In testing MsTMIP test document indexing, we've had some documents that fail to index because of a mismatch between the beginDate value being added, and existing beginDate values already in Solr (same goes for endDate. The core of the issue is that we are using date formats like:

yyyy-mm-ddTHH:MM:ss.SSSZ

Solr's native format is:

yyyy-mm-ddTHH:MM:ssZ

although the Solr schema states that the (optional) milliseconds are allowed.

It looks to me like Solr is reverting to it's native format with any date string values with zeros as milliseconds, like:

1900-01-01T00:00:00.000Z

When this happens, the Solr indexed value is:

1900-01-01T00:00:00Z

When subprocessors (such as the AnnotatorSubprocessor) attempt to merge these fields, the values they are trying to insert (with milli precision) don't match the existing values (seconds precision), and it attempts to add it. In doing so, it attempts to add multiple values to a non-multivalued field, and the insertion fails.
As a temporary hack, I've updated the AnnotatorSubprocessor to compare these fields as Date objects instead of strings. It's not a long term solution, but may fix our immediate issues. Need to discuss with Ben.

#5 - 2015-08-13 23:39 - Ben Leinfelder

Seems like a fine fix to compare Dates instead of strings as long as SOLR isn't discarding/rounding any non-000 millisecond values.

#6 - 2015-10-06 19:22 - Skye Roseboom

Solr seems to not allow trailing 0 in the millisecond value and is intentionally truncating:

http://stackoverflow.com/questions/18920750/solr-triedatefield-changing-the-format-of-date-when-date-ends-with-000-mil-secon

http://lucene.apache.org/solr/5_2_1/solr-core/org/apache/solr/schema/TrieDateField.html

#7 - 2017-03-28 16:16 - Dave Vieglais

- Project changed from CN Index to Infrastructure

- Category changed from d1_cn_index_processor to d1_indexer

- Target version changed from CCI-2.0.0 to CCI-2.4.0

- Milestone set to None