

Infrastructure - Task #8755

Task # 8753 (New): Add support for EML 2.2 (indexing, view)

Expand EML indexing support for EML 2.2

2018-12-19 00:55 - Bryce Mecum

Status:	New	Start date:	2018-12-19
Priority:	Normal	Due date:	
Assignee:	Bryce Mecum	% Done:	0%
Category:		Estimated time:	0.00 hour
Target version:		Story Points:	
Milestone:	None		
Product Version:	*		
Description			

History

#1 - 2019-08-02 21:26 - Bryce Mecum

I'm proposing two changes to the index (code's done, could use a go-ahead):

Overhaul old Annotator code and add a separate EMLAnnotationSubprocessor

The old Annotator subprocessor basically creates a Jena OntologyModel, loads various ontologies into it, and runs SPARQL queries to perform query expansion before populating the sem_annotaton (and other) Solr fields. I copied that code and heavily modified some parts of it to make it faster and more secure. It's faster now because a set of pre-defined ontologies are loaded during init and it's more secure because now other ontologies aren't loaded on demand over the internet.

Add fields to the schema

EML 2.2.0 adds a fair bit of stuff (see https://github.com/NCEAS/eml/blob/BRANCH_EML_2_2/docs/eml-220info.md) but I don't think everything needs to go in the index at this point. I propose we just add semantic annotations and structured funding information at this point.

For semantic annotations, I propose we re-use the sem_annotation field we used for external annotations so the two types can be queried against a single field.

For structured funding, I propose we go with with Lauren did over on metacat-index which is

```
<field name="funding" type="string" indexed="true" stored="true" multiValued="true" />
<field name="fundingText" type="text_general" indexed="true" stored="false" multiValued="true" />
<field name="funderName" type="string" indexed="true" stored="true" multiValued="true" />
<field name="funderIdentifier" type="string" indexed="true" stored="true" multiValued="true" />
<field name="awardNumber" type="string" indexed="true" stored="true" multiValued="true" />
<field name="awardTitle" type="string" indexed="true" stored="true" multiValued="true" />
```

with a copy field to capture fundingText.

#2 - 2019-08-15 01:50 - Bryce Mecum

Need to talk with Dave about integrating this into the next CCI release along with Rob's indexing (ORE) refactor. Warrants some testing.