

Infrastructure - Feature #8053

add funding award number to index

2017-03-28 01:21 - Matthew Jones

|   |                  |                        |            |
|---|------------------|------------------------|------------|
| <b>Status:</b>  | New              | <b>Start date:</b>     | 2017-03-28 |
| <b>Priority:</b>  | Normal           | <b>Due date:</b>       |            |
| <b>Assignee:</b>  | Peter Slaughter  | <b>% Done:</b>         | 0%         |
| <b>Category:</b>  | dataone-cn-index | <b>Estimated time:</b> | 0.00 hour  |
| <b>Target version:</b>  | CCI-2.4.0        | <b>Story Points:</b>   |            |
| <b>Milestone:</b>   | None             |                        |            |
| <b>Product Version:</b>   |                  |                        |            |
| <b>Description</b>  |                  |                        |            |
| Many groups want to track data sets based on the funding award numbers and organizations that funded the work. For example, the Arctic Data Center, BCO-DMO, and R2R all need to report on and search for data based on NSF award numbers. We should add two new multi-valued fields, funding_agency and award_number to the SOLR index so that they can be used for search, display, and faceting. There is a proposal in EML to add structured fields for these, so for details see EML issue <a href="https://github.com/NCEAS/eml/issues/266">https://github.com/NCEAS/eml/issues/266</a> |                  |                        |            |

History

#1 - 2017-03-28 01:21 - Matthew Jones

- Tracker changed from Task to Feature

#2 - 2017-03-28 16:09 - Dave Vieglais

- Project changed from CN Index to Infrastructure
- Category changed from d1\_cn\_index\_common to dataone-cn-index
- Target version set to CCI-2.4.0
- Milestone set to None

#3 - 2019-03-04 19:11 - Bryce Mecum

I'm working on the EML 2.2 indexing stuff and can close this ticket while I do that. To support pre-2.2 docs and 2.2 and onward, I think we should use an XPath of

`//dataset/project/funding/text() | //dataset/project/award/awardNumber/text()`

I think this is the best route to go because it allows us create a single way to search for funding across older and newer EML docs. This is opposed to making clients do queries like `?q:funding:X OR awardNumber: X`. I decided to make the field called funding and define it as a string type with a copyField declaration to fundingText of type text\_general.

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

I decided to do this also to support legacy docs which often stored funding info as free-text strings like "NSF Award Number 123456789".

Do others like this structure and do we also want to support a 'funderName' field for EML2.2 docs?

#### #4 - 2019-03-04 19:36 - Chris Jones

Bryce - Yes, good strategy. One question: Since EML < 2.2.0 documents have `//dataset/project/funding` typed as `EMLTextType`, the call to `text()` there may give back nothing if it's not a text node being returned. This will be the case for the Arctic Data Center, which had to use `//dataset/project/funding/para` to be able to associate multiple funding numbers with a single dataset. I wonder how we can handle the sub-element content of this `EMLText` node. Thoughts?