Infrastructure - Task #8165

Re-factor origin field in isotc211 indexing component

2017-08-29 22:40 - Bryce Mecum

Status:	Closed	**Start date:**	2017-08-29
Priority:	Normal	**Due date:**	
Assignee:	Bryce Mecum	**% Done:**	100%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	CCI-2.4.0		
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

The xpath selectors we use in the origin field in the isotc211 indexing component (bean?) were found to be incorrect for a particular use case and a larger group of us agreed that the usage was incorrect. We should re-visit which xpaths are being used, re-deploy, and re-index the affected content.

I'm pasting in an email chain that initiated the creation of this Issue so we have the full background:

From Chris Turner at Axiom

> Hi Laura and Matt,
>
> Since the launch of the Research Workspace member node, we've no noticed that the dataset citation given at the top of the page doesn't match how we or the PIs would like the official citations to be formatted. There are two issues: selection of contact names for the citation, and appearance of the DOI.
>
> It looks like contact names are being pulled from several parts of the metadata record, the section describing the resource itself (gmd:MD_DataIdentification/gmd:citation/...) and from the section describing associated or aggregated resources (gmd:MD_DataIdentification/gmd:aggregationInfo/...). Here are two examples:
>
> Mary Anne Bishop, Ben Gray, and Scott Pegau. 2017. Fish Predation on Juvenile Herring in Prince William Sound, Alaska, 2009-2012, EVOS Prince William Sound Herring Program. Research Workspace. 10.24431/rw1k1z.
>
> Mary Anne Bishop, Anne Schaefer, Kathy Kuletz, Molly McCammon, Katrina Hoffman, et al. 2017. Fall and Winter Seabird Abundance Data, Prince William Sound, 2007-2017, Gulf Watch Alaska Pelagic Component. Research Workspace. 10.24431/rw1k1w.
>
> In the first exmaple, Scott Pegau is listed in the dataset citation, though in the metadata he is not connected to the dataset but to as the PI for the Herring Program, a larger work referenced in the 'aggregationInfo' section. The second example is the same - McCammon, Hoffman, et al. are pulled from the 'aggregationInfo' element.
>
> Please let me know if I understand correctly how the contacts are being selected for the citation. If I do have it right, is there anything that we can do about it, on our end or in the member node?
>
> The DOI display issue is simpler. DataCite and CrossRef best practices are that DOIs should be displayed as complete URLs, with 'https://doi.org/' appearing before the DOI code assigned to a resource. That's how they're formatted in the metadata records, but the URL-esque formatting is stripped out for the citation and for display in the member node.
>
> Can we display the DOI, both in the citation and in the metadata page as a full link?
>
> Please advise on the best way to remedy these. I apologize if this is not the correct venue for this conversation. Please let me know if it makes more sense to continue talking about this on Slack.
>
> Thanks in advance for your help.
>
> - Chris

From Matt Jones:

> Bryce worked on a new stylesheet for ISO metadata, and so that is scheduled to be released soon. So, any changes would be good to propose even sooner to get them folded in.
>
> If I am interpreting Chris correctly, he's saying that we are indiscriminately pulling Responsible_Party entries regardless of their context in the document, and that it would be best practice to only cite the ResponsibleParty instances that are part of the citatiion in gmd:MD_DataIdentification/gmd:citation/. This makes total sense to me, and is what we do with other metadata standards. So, we need to look into what is causing the behavior, and figure out if it is a stylesheet change, an indexing change, or both.
>
> Matt

From Chris Turner:

> Hi all,
> Matt's understanding is correct. As it is now. ResponsibleParty elements are being pulled in independent of where they appear in the record. Doing as he says and pulling only from gmd:MD_DataIdentification/gmd:citation would go be a good simple fix. We'd also like the pull the contact name from gmd:MD_DataIdentification/gmd:pointOfContact, too.
> I'm available the September 5th-8th to chat if we need to talk that week.
>
> From Laura Moyers:
>
> Thanks, Chris!
>
> Matt, do you think what Chris describes would be a stylesheet change that Bryce could incorporate into his current changes? Would we need to talk with other ISO metadata users about this?
>
> Thanks
> Laura

From Matt Jones:

> I suspect it should be straightforward, and Bryce may have already handled it. We were just talking yesterday about getting his stylesheet changes into a Metacat release and deployed on the CN -- its been in a holding pattern for some reason. So, Jing and Bryce are going to work on getting those display improvements pushed out, as they were requested by a number of members. I don't think Chris' proposal would be at all controversial -- our current display is clearly misleading and would be improved, so I think its a clear win. So. I'll cc Bryce and hopefully he can comment on whether and what work would be needed to support proper citation displays for ISO records.
>
> Matt

From Rob Nahf:

> Will these discussed changes also be reflected in the DataONE solr index? If so, we would likely need to reindex all content of that format after making changes to the parser.
>
> (It sounds like there's broad community support for the change, so reindexing probably wouldn't negatively impact anyone...)

From Bryce Mecum:

> Hey all: Yes, as Matt guesses this is pretty straight forward to fix. The dataset citations in our search and landing pages are powered by our Solr index and we would just need to change the relevant indexing routine and reindex the content. After a quick look at how things are working now, I agree that some change is needed. The information contained in this thread is super helpful so thanks, Chris, for the high level of detail in your original email.

All of that work is on our end and I can coordinate with the DataONE CI team on the changes and we'll let everyone here know when the changes have been made.

The current set of XPaths can be found at https://repository.dataone.org/software/cicore/trunk/cn/d1_cn_index_processor/src/main/resources/application-context-isotc211-base.xml which, at the time of writing this, have this bean set for the origin field:

The sub-tasks here are:

- Figure out what *should* go in there instead
- Probably consult some folks for confirmation
- Update the Bean
- Re-index affected documents once change has been reployed

Related issues:

Related to Infrastructure - Task #8222: reindex all isotc211 content in produ...	**New**	**2017-11-22**

History

#1 - 2017-09-26 00:19 - Bryce Mecum

After looking at this email thread and some example documents, I decided to try simply making the XPath more selective about where in the document it pulls party information. This was suggested in the email thread. I made the change on dev.nceas, uploaded the two example documents linked in the email thread showing the incorrect behavior, reindexed them with the modified indexing bean, and the result was the correct behavior.

The patch is here: https://gist.github.com/amoeba/6670f0ceb8fb1f6cb7fdaf4d46534777

Before this change, the origin field was being filled in with all gmd:CI_ResponsibleParty (individuals and orgs) in the document w/ role originator, author, PI, owner and after the change only those same gmd:CI_ResponsibleParty's under the citation are included.

#2 - 2017-11-21 20:08 - Rob Nahf

- Category set to d1_indexer

- Status changed from New to In Progress

- Assignee set to Bryce Mecum

- Target version set to CCI-2.3.7

- % Done changed from 0 to 30

The unit tests related to the origin field for the isotc211 solr parser are failing, and either the starting science metadata needs to be updated to conform to the tighter restrictions on the origin source fields.

[ERROR] Failures:
[ERROR]
SolrFieldIsotc211Test.testIsotc211DistributionInfoParsing:1126->BaseSolrFieldXPathTest.testXPathParsing:72->BaseSolrFieldXPathTest.compareFields:149 For field: origin
Expected: iterable containing ["NOAA/NESDIS USA, 5200 Auth Rd, Camp Springs, MD, 20746"]
but: item 0: was ""
[ERROR]
SolrFieldIsotc211Test.testIsotc211LooselyCoupledServiceSrvAndDistrib:1138->BaseSolrFieldXPathTest.testXPathParsing:72->BaseSolrFieldXPathTest.compareFields:149 For field: origin
Expected: iterable containing ["Bob"]
but: item 0: was ""

[ERROR]
SolrFieldIsotc211Test.testIsotc211Nodc2FieldParsing:1096->BaseSolrFieldXPathTest.testXPathParsing:72->BaseSolrFieldXPathTest.compareFields:149 For field: origin
Expected: iterable containing ["NEODAAS"]
but: item 0: was ""
[ERROR]
SolrFieldIsotc211Test.testTightlyCoupledServiceSrvOnly:1144->BaseSolrFieldXPathTest.testXPathParsing:72->BaseSolrFieldXPathTest.compareFields:149 For field: origin
Expected: iterable containing ["UNM"]
but: item 0: was ""

corresponding source files are found here:

https://repository.dataone.org/software/cicore/trunk/cn/d1_cn_index_processor/src/test/resources/org/dataone/cn/index/resources/d1_testdocs/isotc211/

#3 - 2017-11-22 18:36 - Bryce Mecum

- Status changed from In Progress to Closed

- % Done changed from 30 to 100

Unit tests all fixed up. It also turned out that the bean XPath (and thereby the unit tests) were wrong because they had hardcoded an element name (gco:CharacterString) where two possible elements are common (gco:CharacterString and gmx:Anchor). I changed the XPath and the tests to reflect the change.

The change will require a reindex of all isotc211 content but this is not a high enough priority to do the reindex now as we're also considering changing this XPath again in https://redmine.dataone.org/issues/8189

#4 - 2017-11-22 19:03 - Rob Nahf

- Related to Task #8222: reindex all isotc211 content in production to reflect final decisions from origin field mapping added

#5 - 2017-11-22 19:09 - Rob Nahf

- Target version changed from CCI-2.3.7 to CCI-2.4.0

Fixes by Bryce made in trunk, so would need to copy to 2.3 branch to pull into earlier releases. It doesn't look like a priority yet, as other tests are failing...