Infrastructure - Bug #8615

isotc211 indexing component has the wrong XPath for the pubDate field

2018-06-14 00:14 - Bryce Mecum

Status:	Closed	**Start date:**	2018-06-14
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

The latest trunk isotc211 indexing component has the following XPATH for pubDate:

```
<bean id="isotc.pubDate" class="org.dataone.cn.indexer.parser.SolrField">
<constructor-arg name="name" value="pubDate"/>
<constructor-arg name="xpath" value="(//gmd:dateStamp/gco:Date/text() | //gmd:dateStamp/gco:DateTime/text())[1]"/>
<property name="converter" ref="dateConverter"/>
</bean>
```

This doesn't make any sense and probably wasn't vetted when it went into version control. If you look at an example document, like one from PANGAEA, you see this is how they describe the publication date:

```
<ns0:identificationInfo>
        <ns0:MD_DataIdentification>
            <ns0:citation>
                <ns0:CI_Citation>
                    <ns0:title>
                        <ns2:CharacterString>Simulated ocean velocity at 420 m water depth for pre-industrial, glacial, Pliocene, and Miocene climate states</ns2:CharacterString>
                    </ns0:title>
                    <ns0:date>
                        <ns0:CI_Date>
                            <ns0:date>
                                <ns2:DateTime>2018-04-25T15:20:13</ns2:DateTime>
                            </ns0:date>
                            <ns0:dateType>
                                <ns0:CI_DateTypeCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_DateTypeCode" codeListValue="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_DateTypeCode_publication">publication</ns0:CI_DateTypeCode>
                            </ns0:dateType>
                        </ns0:CI_Date>
                    </ns0:date>
```

I think it'd be good if the XPath pulled out the first date from the identificationInfo/citation, preferring one of dateType publicationDate and falling back to any date that's in the identificationInfo/citation.

History

#1 - 2019-03-21 20:39 - Roger Dahl

Example:

https://search.dataone.org/view/http://get.iedadata.org/metadata/iso/609441

Publication date in the metadata as 2010, but the pubDate value in Solr is 2018-05-17T00:00:00Z

#2 - 2019-04-01 22:10 - Jing Tao

The xpath looks like:
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:Date/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:Date[1]/text()
|
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:DateTime/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:DateTime[1]/text()

#3 - 2019-04-15 21:17 - Jing Tao

- % Done changed from 0 to 100

- Assignee set to Jing Tao

- Status changed from New to Closed

#4 - 2019-06-01 06:22 - Jing Tao

- % Done changed from 100 to 30

- Status changed from Closed to In Progress

#5 - 2019-06-01 06:25 - Jing Tao

ISO fragment:

```
    <gmd:MD_DataIdentification>
      <gmd:citation>
        <gmd:CI_Citation>
          <gmd:title>
            <gco:CharacterString>Steller sea lion (Eumetopias jubatus) satellite telemetry data used to determ
ine at-sea distribution in the western-central Aleutian Islands, Alaska 2000-2013</gco:CharacterString>
          </gmd:title>
          <gmd:date>
            <gmd:CI_Date>
              <gmd:date>
                <gco:Date>2013-01-01</gco:Date>
              </gmd:date>
              <gmd:dateType>
                <gmd:CI_DateTypeCode codeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schema/resource
s/Codelist/gmxCodelists.xml#CI_DateTypeCode" codeListValue="creation">creation</gmd:CI_DateTypeCode>
              </gmd:dateType>
            </gmd:CI_Date>
          </gmd:date>
          <gmd:date>
            <gmd:CI_Date>
              <gmd:date>
                <gco:Date>2019-06-01</gco:Date>
              </gmd:date>
              <gmd:dateType>
                <gmd:CI_DateTypeCode codeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schema/resource
s/  Codelist/gmxCodelists.xml#CI_DateTypeCode" codeListValue="publication">publication</gmd:CI_DateTypeCode>
              </gmd:dateType>
            </gmd:CI_Date>
          </gmd:date>
```

Our processor will get the first date time 2013-0-01.

#6 - 2019-06-01 06:30 - Jing Tao

Our rules look like:

/gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:Date/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:Date[1]/text()
|
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:DateTime/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:DateTime[1]/text()

If I removed the general one -

//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:Date[1]/text()
We can get the correct result. So the processor doesn't apply the xpath by the order?

#7 - 2019-06-03 22:25 - Bryce Mecum

Hey Jing, is the XPath in your above comment the one in the source? It looks like it has a bug (first element is missing a seoncd /) and also duplicates the XPaths.

Should it be:

```
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:DateTime/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:DateTime[1]/text()
```

#8 - 2019-06-03 22:42 - Jing Tao

Hi Bryce:
The missing / is a typo on the comment. The code doesn't miss it. Sorry to confuse you.

The operator | is to compute two node sets:

https://www.w3schools.com/xml/xpath_operators.asp

So we should use:

```
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:Date/text()
```

```
 |//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:DateTime/text()
```

The above two xpath are not duplicated - the last elements are different.

#9 - 2019-06-03 22:54 - Bryce Mecum

Gotcha, my mistake. The fallback XPath still seems desirable though. What if we just ordered them differently?

```
//gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd:d
ateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:Date/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date[following-sibling::gmd
:dateType/gmd:CI_DateTypeCode/text() = 'publication']/gco:DateTime/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:Date[1]/text()
| //gmd:identificationInfo/*/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date/gco:DateTime[1]/text()
```

#10 - 2019-06-03 23:02 - Jing Tao

To my understand, the operator | is not or. The xpath will get all values of the all existing path. In Roger's case, the xml object has both xpath (fallback one and publication one). So it will get two values. Since the fallback's position in the xml instance is prior to the publication one, the fallback one was selected.
I am not sure how can we keep the fallback path.

#11 - 2019-06-03 23:26 - Bryce Mecum

You are right, | is not \or.

I was thinking that re-arranging the order would any nodeset returned by the XPath ordered from most desirable to least desirable. After a quick test with an example doc I made up it does seem like the | doesn't return an ordered nodeset which was my hope. Or if it is ordered, it's not ordered in the same order as the XPath.

From the help https://stackoverflow.com/questions/5497197/how-to-get-the-real-node-order-from-xpath-expression-java it looks like the XPath spec defines nodesets as unordered and implementations of the spec may choose to allow clients to enforce an order. Do you know if we can do that here?

#12 - 2019-06-03 23:29 - Bryce Mecum

Looking at that a bit more, maybe what's really going on is that the nodeset is returned in document order rather than XPath order which matches what I'm seeing after doing some testing.

#13 - 2019-06-03 23:32 - Jing Tao

Yes, I think they are returned in the document order.

#14 - 2019-07-02 22:22 - Jing Tao

- Status changed from In Progress to Closed

- % Done changed from 30 to 100

Now we used Saxon to support xpath 2.0 and fixed the issue.