



Primo File Splitters

Tony Gibbons

2012 February 9

Copyright Statement

All of the information and material inclusive of text, images, logos, product names is either the property of, or used with permission by Ex Libris Ltd. The information may not be distributed, modified, displayed, reproduced – in whole or in part – without the prior written permission of Ex Libris Ltd.

TRADEMARKS

Ex Libris, the Ex Libris logo, Aleph, SFX, SFXIT, MetaLib, DigiTool, Verde, Primo, Voyager, MetaSearch, MetaIndex and other Ex Libris products and services referenced herein are trademarks of Ex Libris, and may be registered in certain jurisdictions. All other product names, company names, marks and logos referenced may be trademarks of their respective owners.

DISCLAIMER

The information contained in this document is compiled from various sources and provided on an "AS IS" basis for general information purposes only without any representations, conditions or warranties whether express or implied, including any implied warranties of satisfactory quality, completeness, accuracy or fitness for a particular purpose.

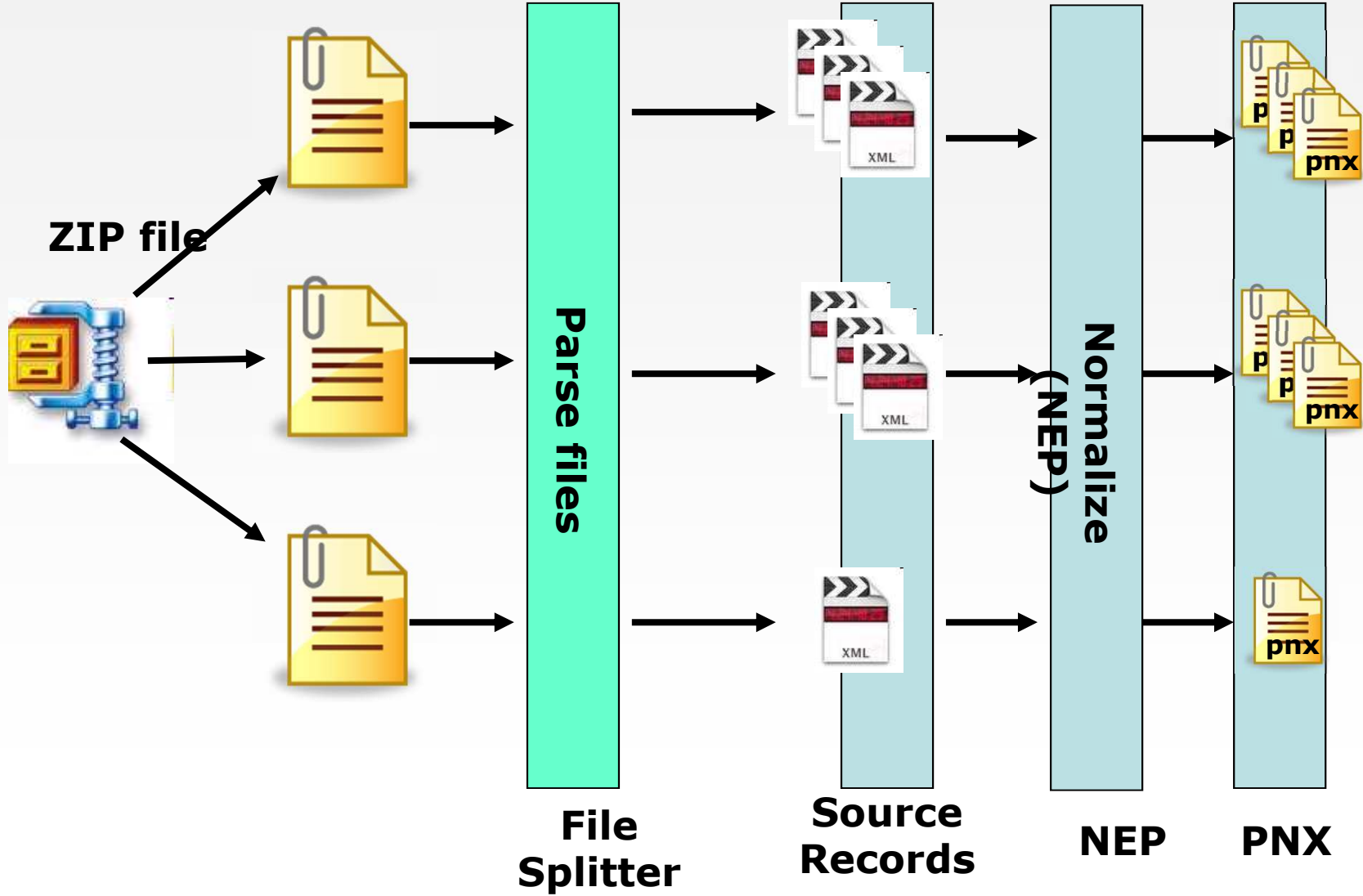
Ex Libris, its subsidiaries and related corporations ("Ex Libris Group") disclaim any and all liability for all use of this information, including losses, damages, claims or expenses any person may incur as a result of the use of this information, even if advised of the possibility of such loss or damage.

© Ex Libris Ltd., 2012

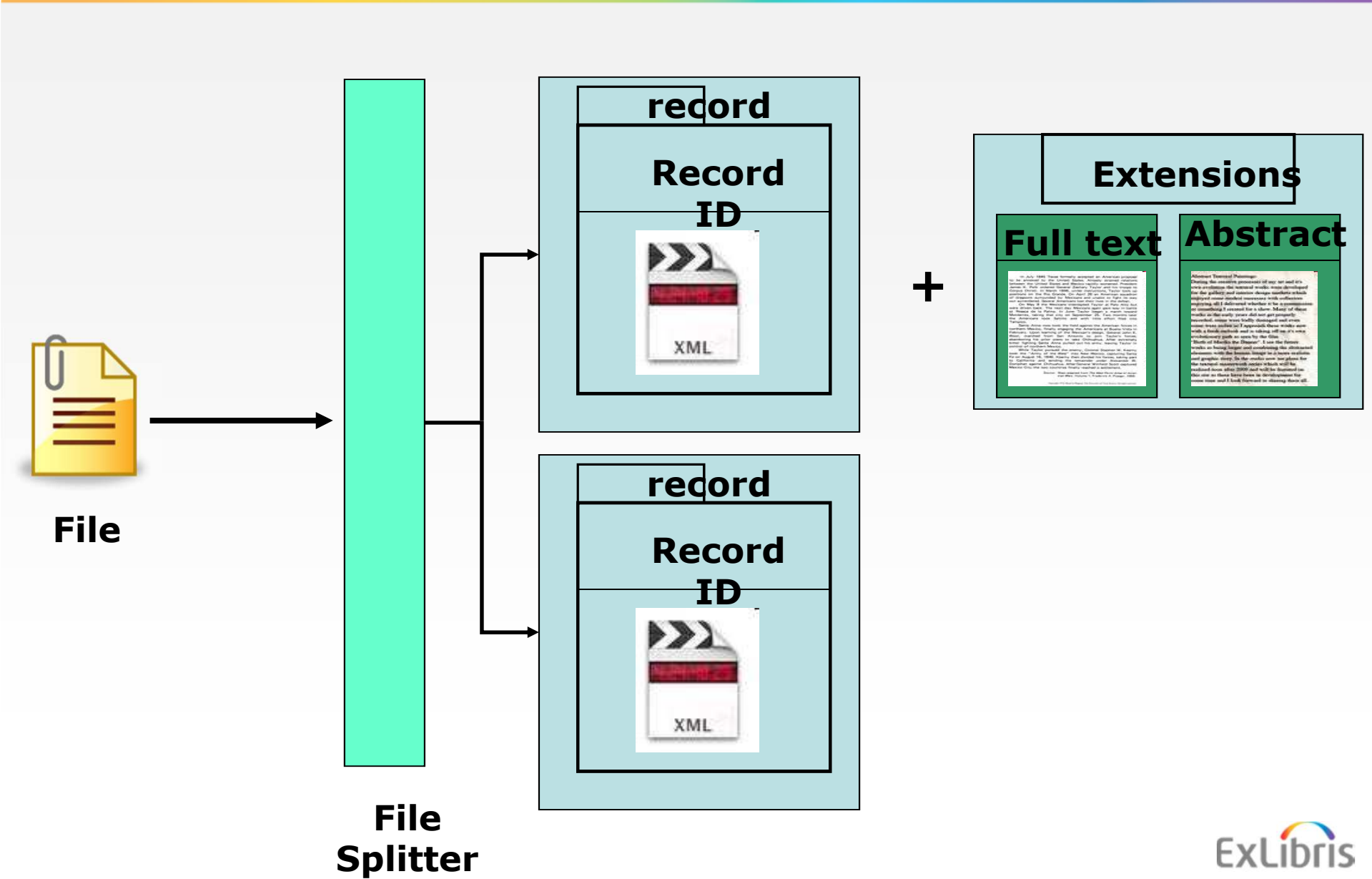
File splitters (Harvesting)

- Primo V2 supports only 2 file types:
 - XML in OAI structure
 - MARC Exchange
1. What if I want to harvest an **XML** file **not** in an **OAI** format? (Use **XSLT**?)
 2. What if I want to harvest an **Excel** file?
 3. How do I harvest "**Primo Central**" data without file splitters?

File splitters



File splitters (input/output)



File splitters

- A file splitter is associated to a data source.

> **Data Sources**

Data Source Attributes for PRIMO-ALEPH

Source Description	Source name: <input type="text" value="PRIMO-ALEPH"/>	Source code: primo_aleph
	Description: <input type="text" value="Reference data"/>	
Source Definition	Institution: <input type="text" value="Primo Institution"/>	Source system: <input type="text" value="Aleph"/>
	File Splitter: <input type="text" value="OAI splitter"/> demoXMLSplitter test OAI splitter Static OAI splitter MARC Exchange splitter SFXXML splitter WARC splitter MyXMLSplitter JorgenFS Excel Splitter	Character Set: <input type="text" value="UTF-8"/>
		Transformation file name <input type="text"/>

Created Jan 19, 201 By primo

File splitters

“**Static**” Out of the Box

- OAI
- MARC
- SFX
- WARC

Configurable Out of the Box

- Generic XML
- Generic HTML

Implement your own specific splitter

- JAVA coding

Register Your XML File Splitter

Primo Back Office:

Advanced Configuration / All Mapping Tables / SubSystem:
Publishing / Table Name: File Splitters

> Mapping Tables

Sub System : Publishing

Table Name : File Splitters

Mapping Table Rows

Enabled	Name	Splitter Class
<input type="checkbox"/>	<input type="text"/>	<input type="text"/>
<input checked="" type="checkbox"/>	demoXMLSplitter	generic.DomXmlSplitter
<input checked="" type="checkbox"/>	test	generic.DomXmlSplitter
<input checked="" type="checkbox"/>	OAI splitter	splitters.oai.OAISplitterCB
<input checked="" type="checkbox"/>	Static OAI splitter	splitters.oai.OAISplitterCB
<input checked="" type="checkbox"/>	MARC Exchange splitter	marc_exchange.MarcExchangeSplitter
<input checked="" type="checkbox"/>	SFX XML splitter	generic.DomXmlSplitter
<input checked="" type="checkbox"/>	WARC splitter	splitters.warc.WarcSplitter
<input checked="" type="checkbox"/>	MyXMLSplitter	generic.DomXmlSplitter
<input checked="" type="checkbox"/>	JorgenFS	generic.DomXmlSplitter
<input checked="" type="checkbox"/>	Excel Splitter	demo.ExcelFileSplitter

XML file splitter (Parameters)

Primo Back Office:

Advanced Configuration / All Mapping Tables / SubSystem:
Publishing / Table Name: File Splitters Params

Sub System : Table Name :

Mapping Table Rows

Enabled	Param Name	Param Value	File Splitter name
<input type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="Select Value"/>
<input checked="" type="checkbox"/>	StatusXPath	/OAI-PMH/ListRecords/record/hea	OAI splitter
<input checked="" type="checkbox"/>	IdentifierXPath	/OAI-PMH/ListRecords/record/hea	OAI splitter
<input checked="" type="checkbox"/>	ContentXPath	/OAI-PMH/ListRecords/record/met	OAI splitter
<input checked="" type="checkbox"/>	SplitByXPath	/OAI-PMH/ListRecords/record	OAI splitter
<input checked="" type="checkbox"/>	StatusXPath	/Repository/ListRecords/record/he	Static OAI splitter
<input checked="" type="checkbox"/>	IdentifierXPath	/Repository/ListRecords/record/he	Static OAI splitter
<input checked="" type="checkbox"/>	ContentXPath	/Repository/ListRecords/record/mi	Static OAI splitter
<input checked="" type="checkbox"/>	SplitByXPath	/Repository/ListRecords/record	Static OAI splitter
<input checked="" type="checkbox"/>	RootXPath	collection	SFXXML splitter

Mandatory Parameters (XML)

Name	Description	Explanation
RootXPath	The tag that brackets the data elements	<pre data-bbox="1265 359 1814 526"><record> <metadata> </metadata> <field1> </field1> </record></pre> <p data-bbox="1265 574 1702 614">Use the value: record</p>
FullRecordXPath	The beginning of a record . A file may contain more than one record	<pre data-bbox="1265 657 1870 825"><record> <metadata> </metadata> <field1> </field1> </record></pre> <p data-bbox="1265 873 1915 912">Use the value: record/metadata</p>
IdentifierXPath	The XPath to the identifier tag of the record.	<pre data-bbox="1265 944 1870 1112"><record> <metadata> </metadata> <ID> </ID> </record></pre> <p data-bbox="1265 1160 1691 1248">Use the value: record/metadata/id</p>

Additional Parameters (XML)

Name	Description
StatusXPath + StatusWhenDeleted	<pre data-bbox="1037 379 1478 638"><records> <record> <id>123</id> <status>Y</status> </record> </records></pre> <p data-bbox="1037 691 1626 770">StatusXPath: //record/status StatusWhenDeleted = Y</p>
ExternalResourceSourceXPath	<pre data-bbox="1037 895 1749 1106"><records> <record> <path>www.xxx.com/file.pdf</path> </record> </records></pre> <p data-bbox="1037 1118 1615 1198">ExternalResourceSourceXPath: //record/path</p>
AddExtensionsToExtensionsTable	

HTML file splitter

Class:

com.exlibris.primo.publish.platform.harvest.splitters.html.HTMLFileSplitter

The resulting XML:

```
<record>
```

```
  <description>Free Web tutorials</description>
```

Meta tags

```
  <keywords> HTML,CSS,XML,JavaScript</keywords>
```

```
  <author>Hege Refsnes</author>
```

```
  <substrings>
```

HTML fragment

```
    <sub1>blah blah blah...</sub1>
```

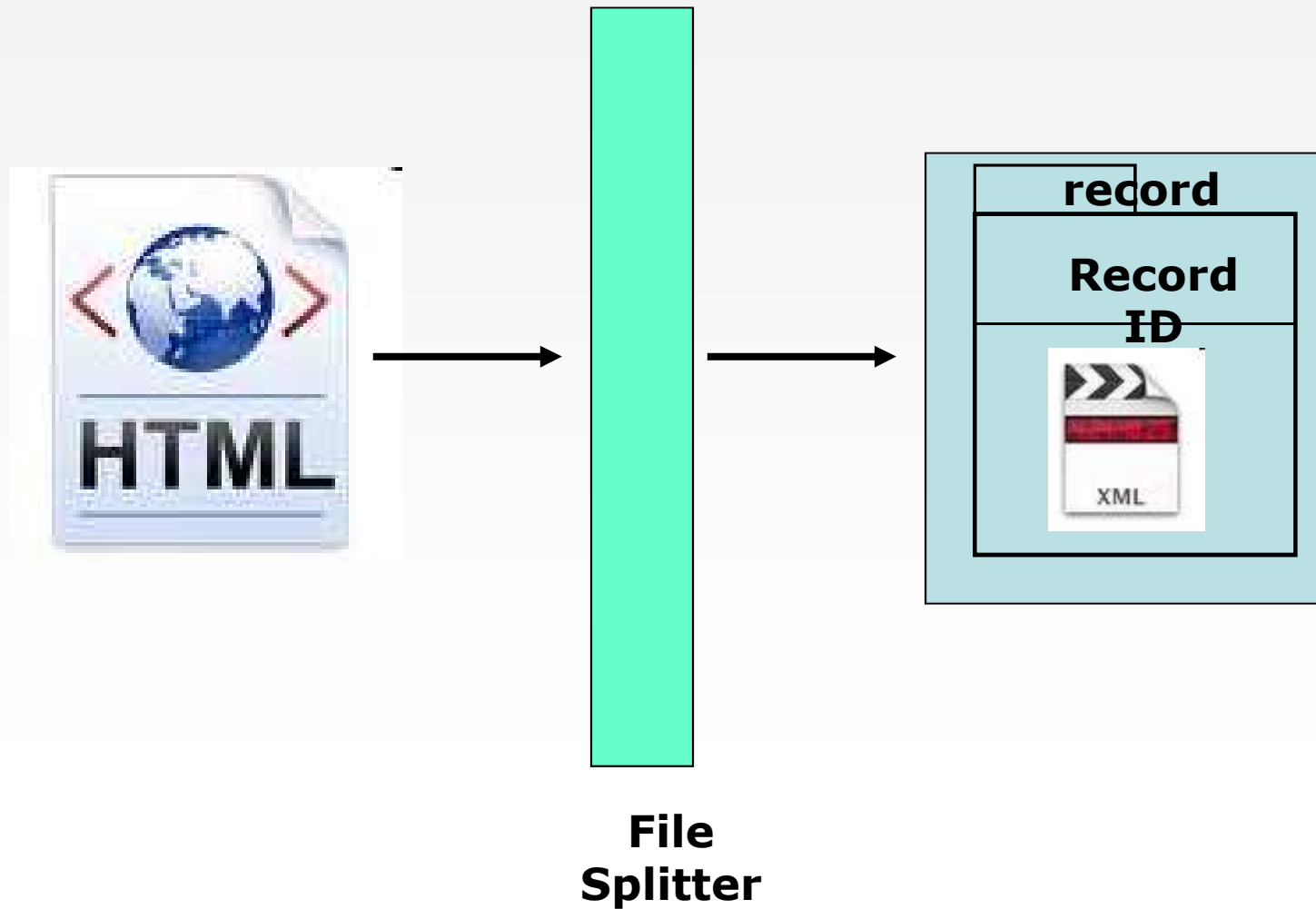
```
    <sub2>kuku kuku kuku...</sub2>
```

```
  </substrings>
```

```
</record>
```

HTML file splitter

- HTML file = 1 record (not more)



Mandatory Parameter HTML

Name	Description	Explanation	Mandatory
IdentifierXpath	A meta param name holding the unique identifier of the record.	For example for the following file: <pre data-bbox="1160 639 1666 986"><html> <head> <meta name="id">123</meta> </head> <body> </body> </html></pre> IdentifierXpath = id	Yes

Additional Parameters HTML

Name	Description	Explanation
StatusXpath + StatusWhenDeleted	For example for the following file: <pre data-bbox="824 363 1480 632"><html> <head> <meta name="status">Y</meta> </head> <body> </body> </html></pre> <p data-bbox="824 675 1249 743">IdentifierXpath = status StatusWhenDeleted = Y</p>	
ExternalResourceSourceXpath	<pre data-bbox="824 850 1290 1190"><html> <head> <meta name="path"> www.xxx.com/file.pdf </meta> </head> <body> </body> </html></pre> <p data-bbox="824 1241 1464 1278">ExternalResourceSourceXpath: path</p>	Same as for the XML file splitter except this is a name of a meta param and not an XPath.
AddExtensionsToExtensionsTable		If not set, will be under the <substrings> section

File splitters (indexing extensions)

- Must make sure 2 settings are set!!!



- No settings = **NO indexing**

File splitters indexing extensions

- Default type = FULLTEXT

> Mapping Tables

Sub System : Table Name :

Mapping Table Rows

Enabled	EXTENSION NAME	EXTENSION VALUE	Target Tag Path
<input type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="checkbox"/>	REVIEW	<input type="text"/>	search:review
<input checked="" type="checkbox"/>	FULLTEXT	<input type="text"/>	search:fulltext
<input checked="" type="checkbox"/>	TAG	value	search:usertag
<input checked="" type="checkbox"/>	POPULARITY	value	sort:popularity
<input type="checkbox"/>	toc	<input type="text"/>	search:toc
<input type="checkbox"/>	abstract	<input type="text"/>	search:abstract
<input type="checkbox"/>	fiction	<input type="text"/>	search:fiction
<input checked="" type="checkbox"/>	TOC_BT	<input type="text"/>	search:toc

NOTE: Must enable the row!!!

File splitters indexing extensions

- Mark PNX with extensions

> Mapping Tables

Sub System : Publishing

Table Name : Datasource Index Extensions

Mapping Table Rows

Enabled

Data Source Name

Type

Select Value

Select Value



WARC

Index All

Index All

Index If Exists

Table Description: Determine if need to index extensions

Documentation

EL Commons

<http://www.exlibrisgroup.org/display/PrimoOI/File+Splitter+Plug-In>



Hello Tony Gibbons ([Log Out](#))

[click here for EL Commons Wiki](#)

You are here: [EL Commons](#) > [CodeShare](#) > [Primo Open Interfaces](#) > [Plug-Ins](#) > [File Splitter Plug-In](#)

[My Area](#) [View](#) [Edit](#) [Aleph](#) [Voyager](#) [Primo](#) [MetaLib](#) [SFX](#) [bX](#) [Verde](#) [Rosetta](#) [DigiTool](#) [Cross Product](#) [Presentations](#) [RSS feed builder](#)

[Files \(23\)](#)

Labels: [plug-in](#) [primo](#)

File Splitter Plug-In

Thank You!

Tony Gibbons, MLIS

tony.gibbons@exlibrisgroup.com

File splitter (Plug-in)

Flow:

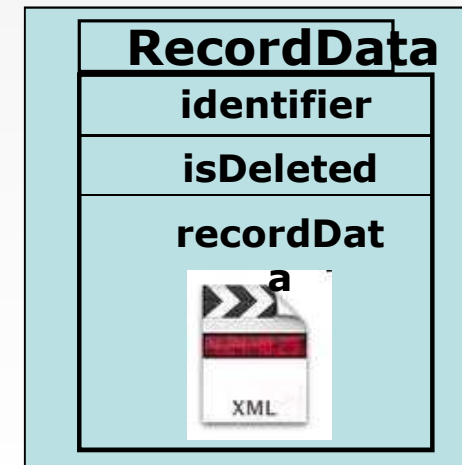
- Copy a JAR file from Primo into your environment
- Implement your Plug-in's
- Copy back the implementations into Primo

File splitter (Plug-in)

RecordData –

An object representing a parsed record

- **identifier** –unique record id.
- **recordData** –XML file.
- **isDeleted** - is for deletion

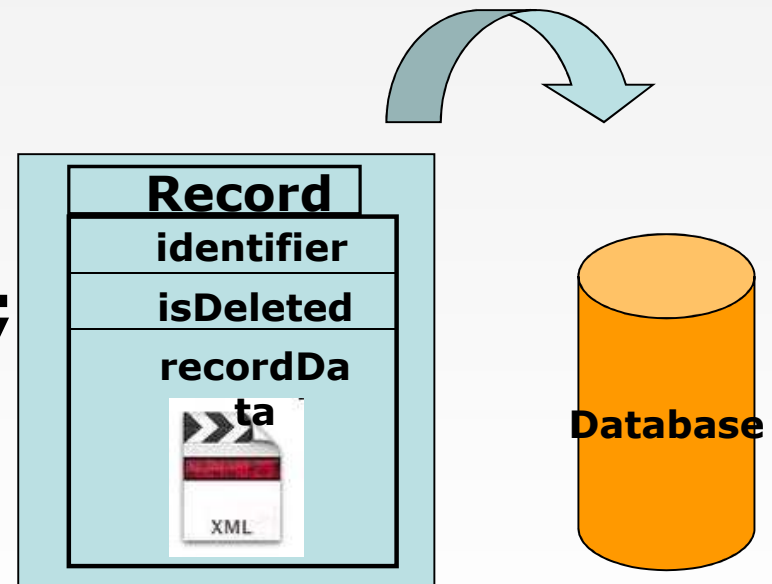


File splitter (Plug-in)

IRecordSaver –

Knows to save *RecordData* objects.

```
RecordData record =  
getNextRecord();  
recordSaver.save(record);
```



File splitter (Plug-in)

IFileSplitter :

#1: init()

1. **charSet** –
2. **logger** – write to log.
3. **Params** – “File Splitters Params”.

> Data Sources

Data Source Attributes for PRIMO-ALEPH

Source Description

Source name:

PRIMO-ALEPH

Source code:

primo_aleph

Description:

Reference data

Source Definition

Institution:

Primo Institution

Source system:

Aleph

File Splitter:

OAI splitter

The character set

Character Set:

UTF-8

File splitter (Plug-in)

IFileSplitter :

#2: parse()

1. InputStream –

2. IRecordSaver – Saves RecordData



File splitter (Plug-in)

IFileSplitter :

#3: doneParsing()

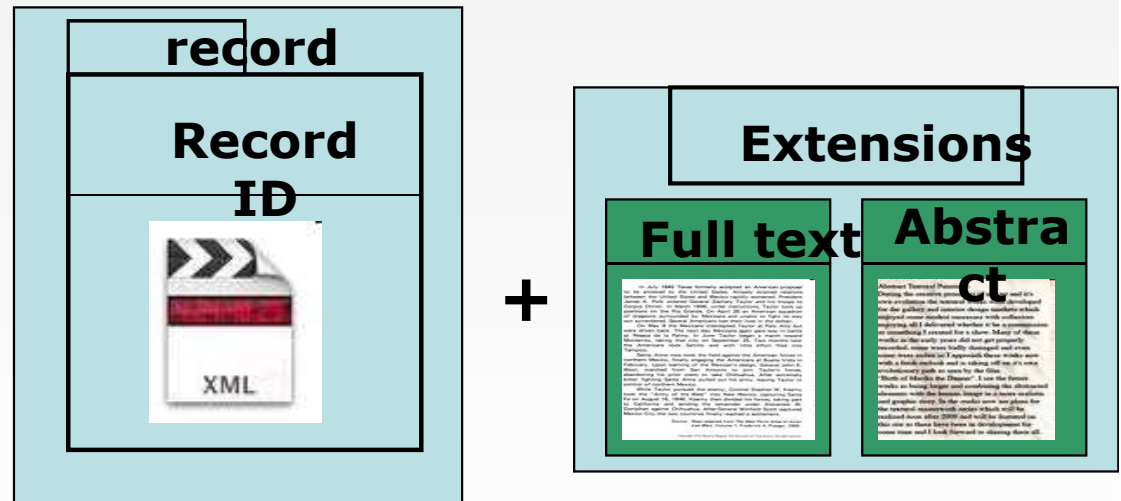
Called when no more files are left for parsing.

File splitter (Plug-in)

ExtensionData:

Represents an extension

1. Type
2. Value
3. Overwrite



```
RecordData record = getNextRecord();
```

```
record.addExtension(new ExtensionData("FULLTEXT", "blah blah balh..."));  
record.addExtension(new ExtensionData("ABSTRACT", "kuku kuku  
kuku..."));
```

```
recordSaver.save(record);
```

File splitter (Plug-in)

IRecordSaverWrapper-

Have control on the OTB file splitters at the end of parsing.

1: doBeforeSave()

What can be done:

1. Decide to skip record
2. Scan info from all records – pass it to the EndHandler

```
1. RecordData
   Boolean save =
   saverWrapper.doBeforeSave
   ve(rd)

   if (save) {
   saveRecord(rd);
   }
```

Configure in the FS parameters

Parameter name:

RecordSaverWrapper

Parameter value: A class name

File splitter (Plug-in)

IEndParsingHandler -

Have control on the OTB file splitters at the end of parsing.

#1: `init()`

1. `RecordSaver`

#2 `execute()`

`splitter.doneParsing();`

`Handler.execute();`

`recordSaver.saveLeftOvers(config);`

What can be done:

1. Add additional records, not existing in harvested files.
2. Process data collected by the `IRecordSaverWrapper`

Configure in the FS parameters

Parameter name: EndParserHandler

Parameter value: A class name