Primo File Splitters

Tony Gibbons

2012 February 9

Copyright Statement

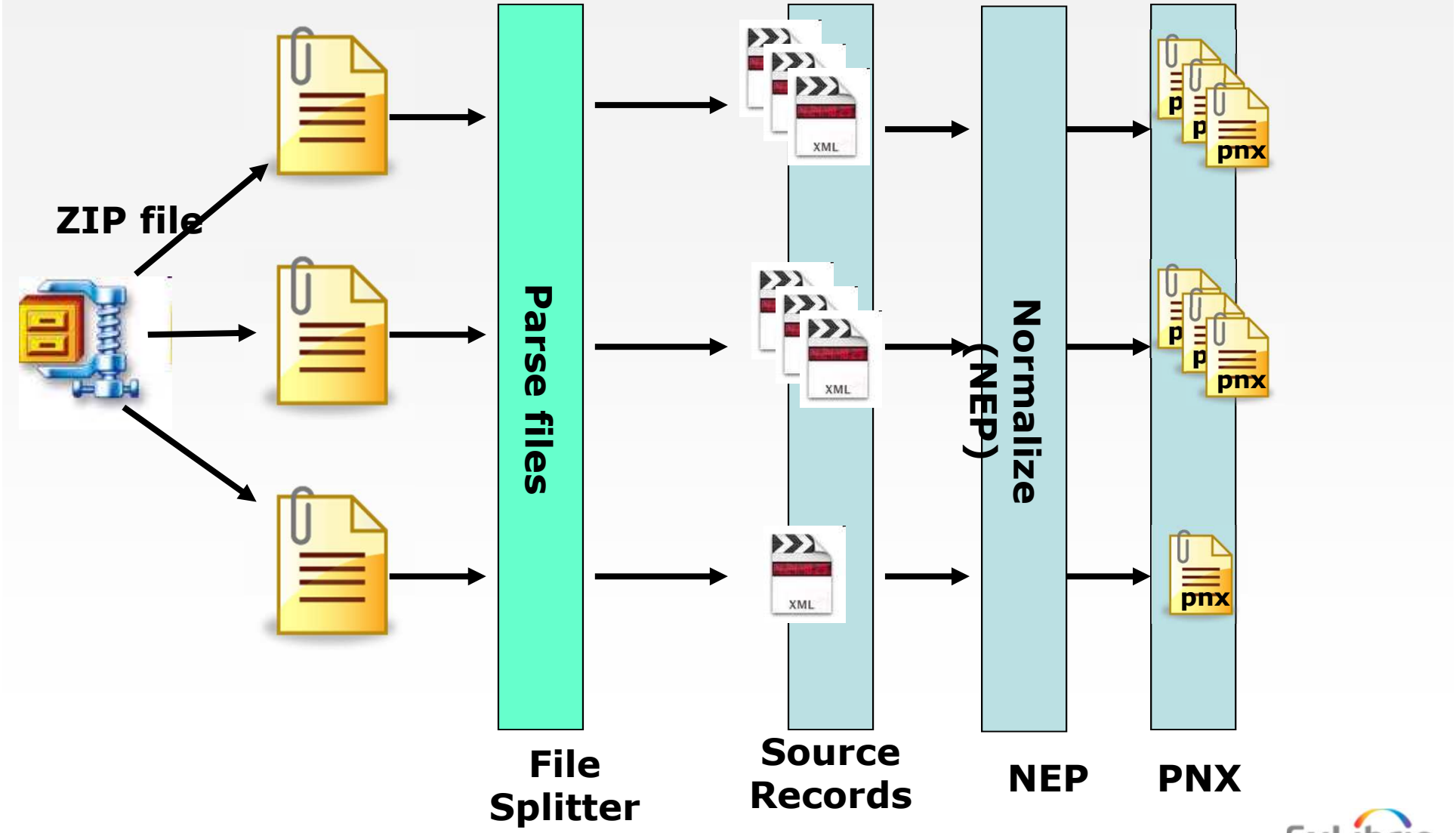
ExLibris

File splitters (Harvesting)

- Primo V2 supports only 2 file types:
  - XML in OAI structure
  - MARC Exchange

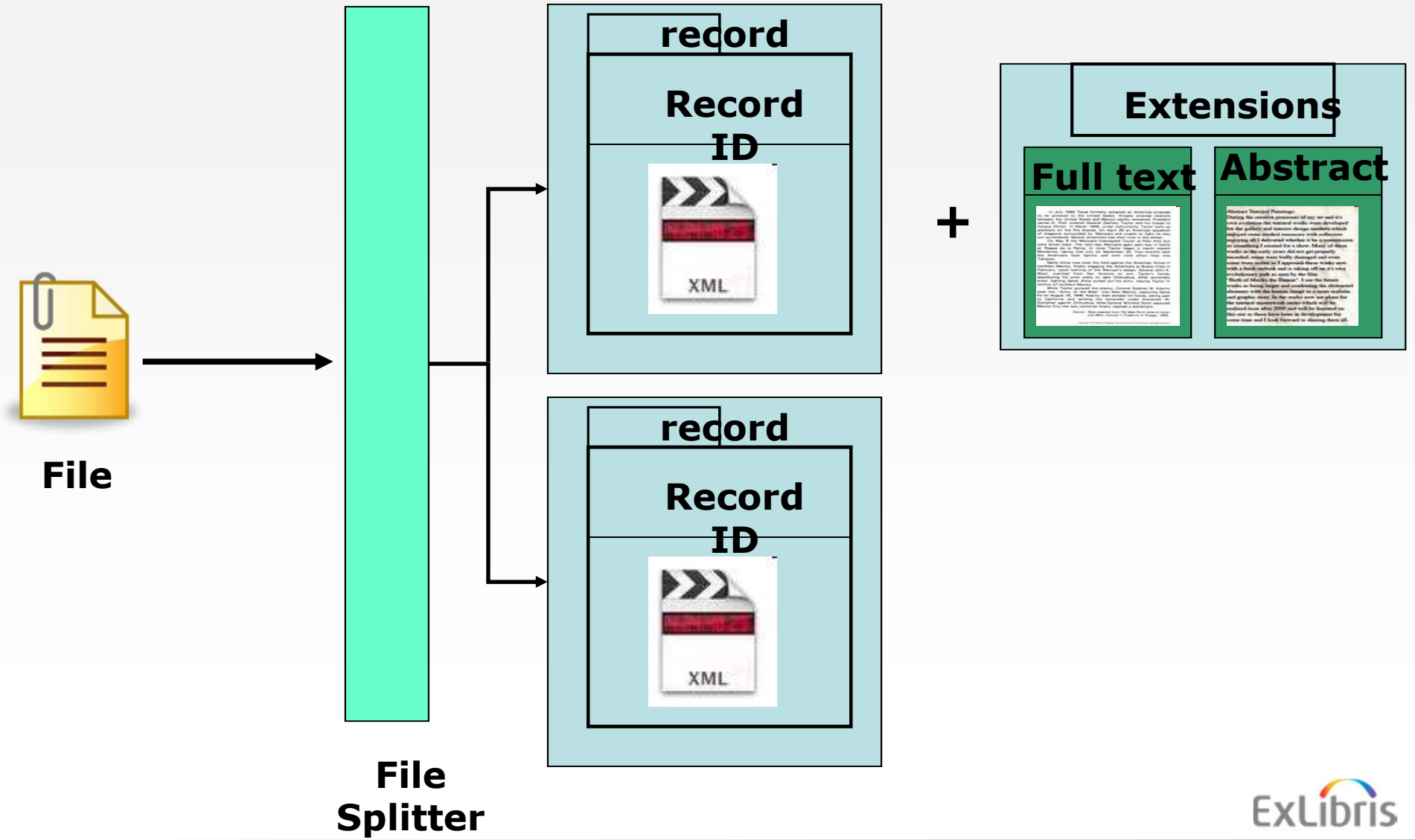
1. What if I want to harvest an **XML** file **not** in an **OAI** format? (Use **XSLT**?)
2. What if I want to harvest an **Excel** file?
3. How do I harvest **"Primo Central"** data without file splitters?

File splitters



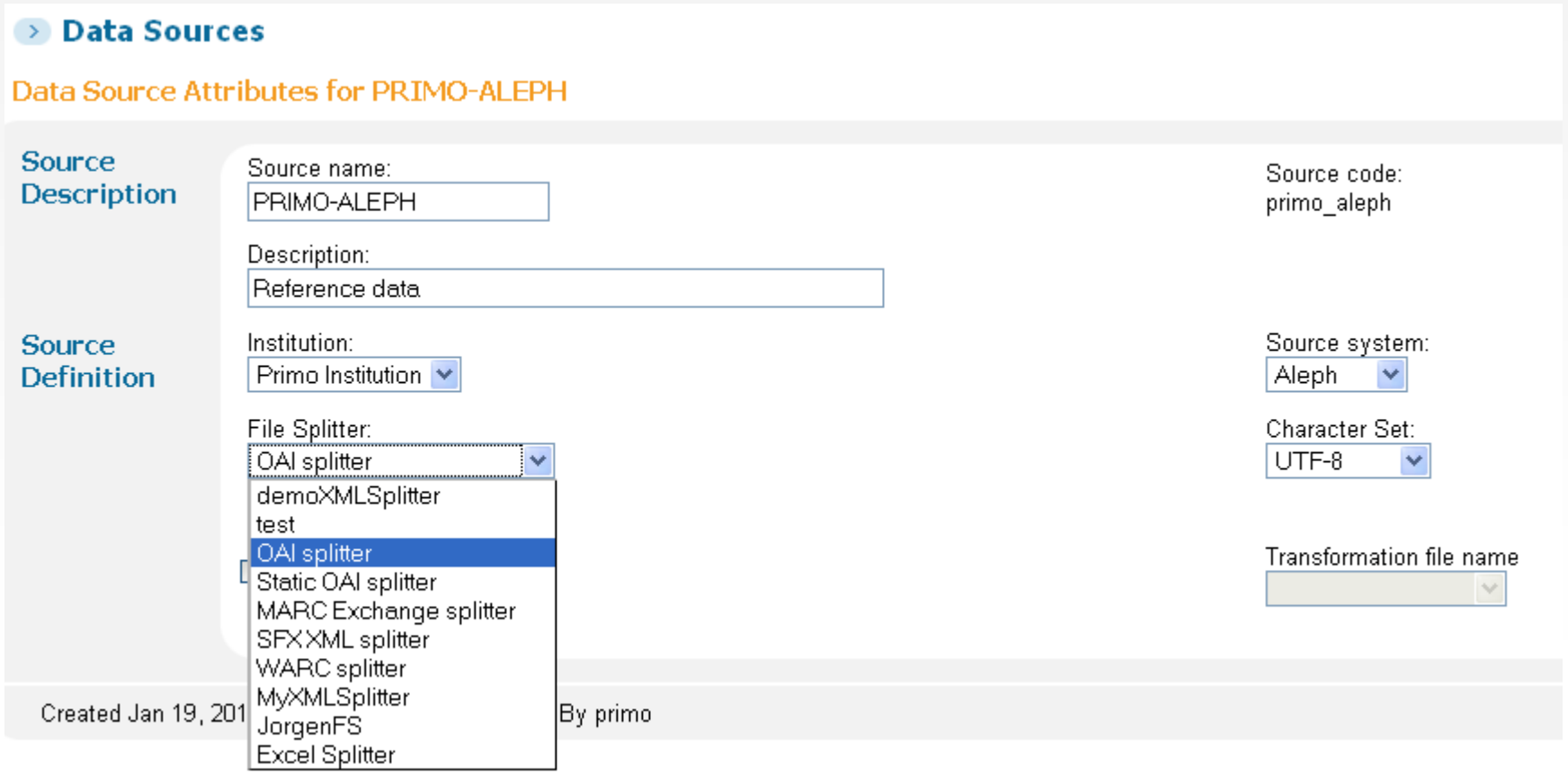
ExLibris

File splitters (input/output)



File splitters

- A file splitter is associated to a data source.



File splitters

"**Static**" Out of the Box

- OAI
- MARC
- SFX
- WARC

Configurable Out of the Box

- Generic XML
- Generic HTML

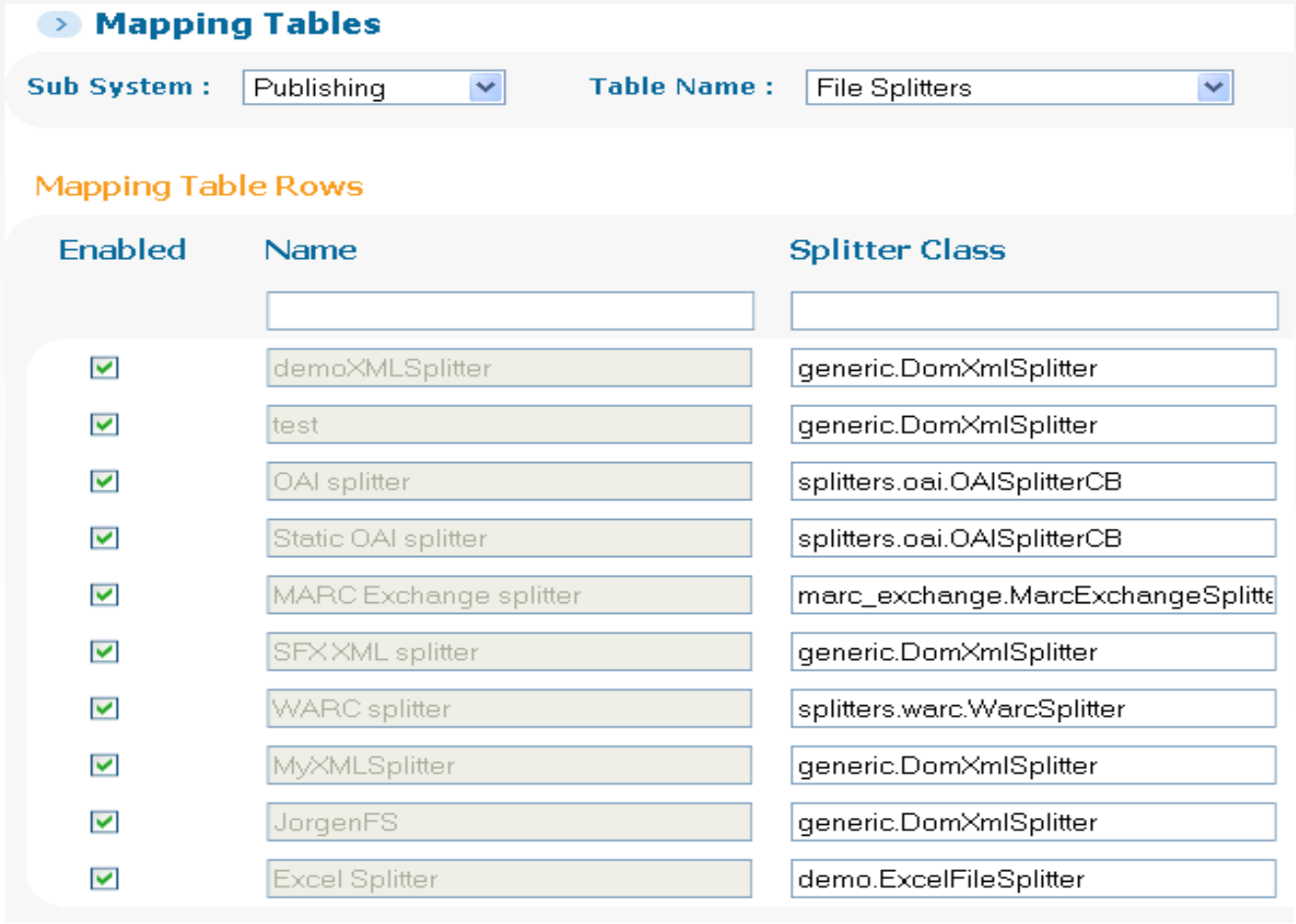
Implement your own specific splitter

- JAVA coding

Register Your XML File Splitter

Primo Back Office:

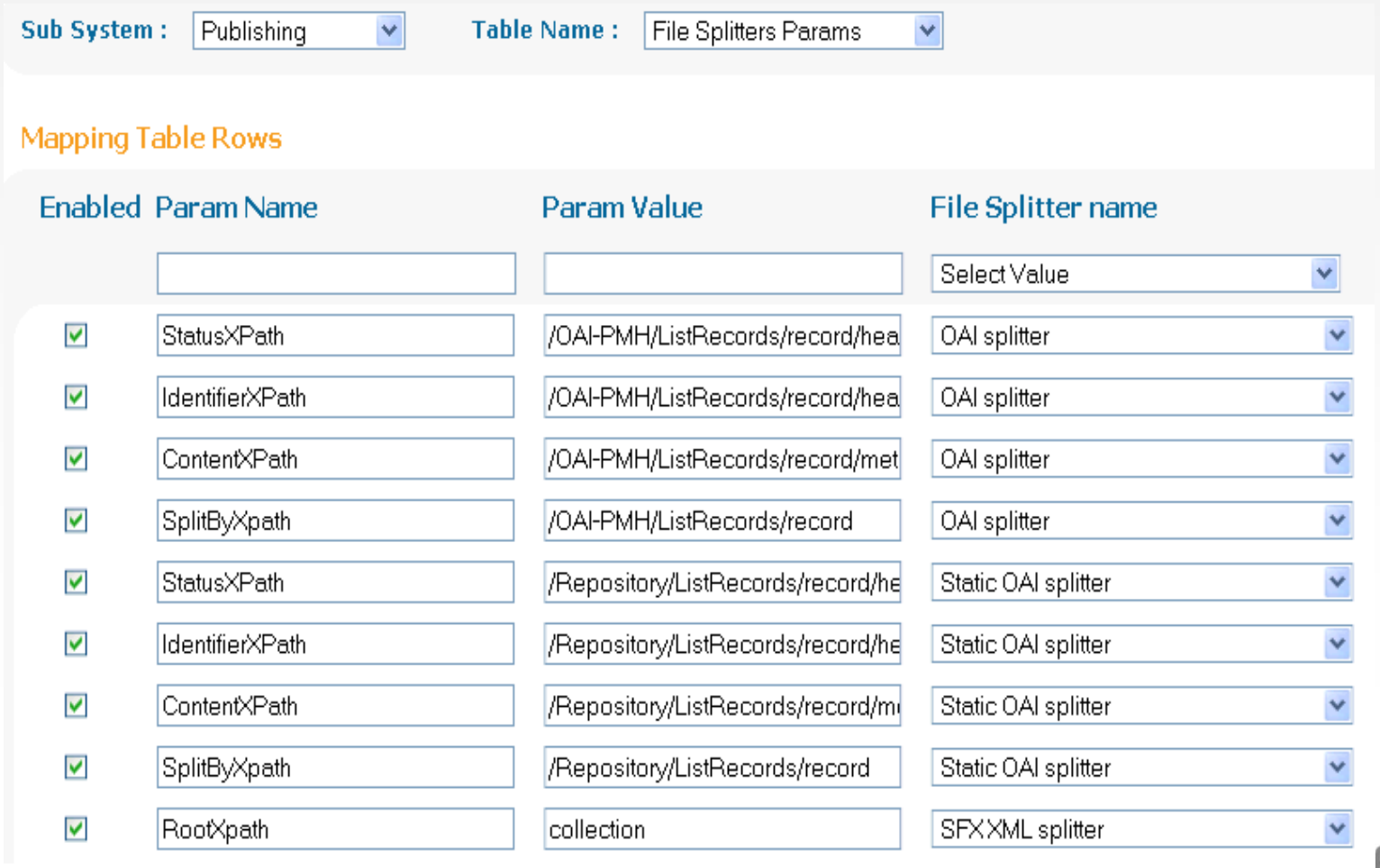
Advanced Configuration / All Mapping Tables / SubSystem: Publishing / Table Name: File Splitters



XML file splitter (Parameters)

Primo Back Office:

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



Mandatory Parameters (XML)

Name	Description	Explanation
RootXpath	The **tag that brackets the data elements**	<**record**> <metadata> </metadata> <field1> </field1> </**record**> Use the value: **record**
FullRecordXPath	The **beginning of a record**. A file may contain more than one record	<**record**> **<metadata> </metadata>** <field1> </field1> </**record**> Use the value: **record/metadata**
IdentifierXpath	The XPath to the **identifier tag** of the record.	<**record**> **<metadata> </metadata>** <ID> </ID> </**record**> Use the value: **record/metadata/id**

ExLibris

Additional Parameters (XML)

Name	Description
StatusXpath + StatusWhenDeleted	<records> <record> <id>123</id> <status>Y</status> </record> </records> StatusXpath:**//record/status** StatusWhenDeleted **= Y**
ExternalResourceSourceXpath	<records> <record> <path>www.xxx.com/file.pdf</path> </record> </records> ExternalResourceSourceXpath: **//record/path**
AddExtensionsToExtensionsTable	

ExLibris

HTML file splitter

Class:

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

The resulting XML:

```
<record>
    <description>Free Web tutorials</description>
    <keywords> HTML,CSS,XML,JavaScript</keywords>
    <author>Hege Refsnes</author>
    <substrings>
        <sub1>blah blah blah…</sub1>
        <sub2>kuku kuku kuku…</sub2>
    </substrings>
</record>
```

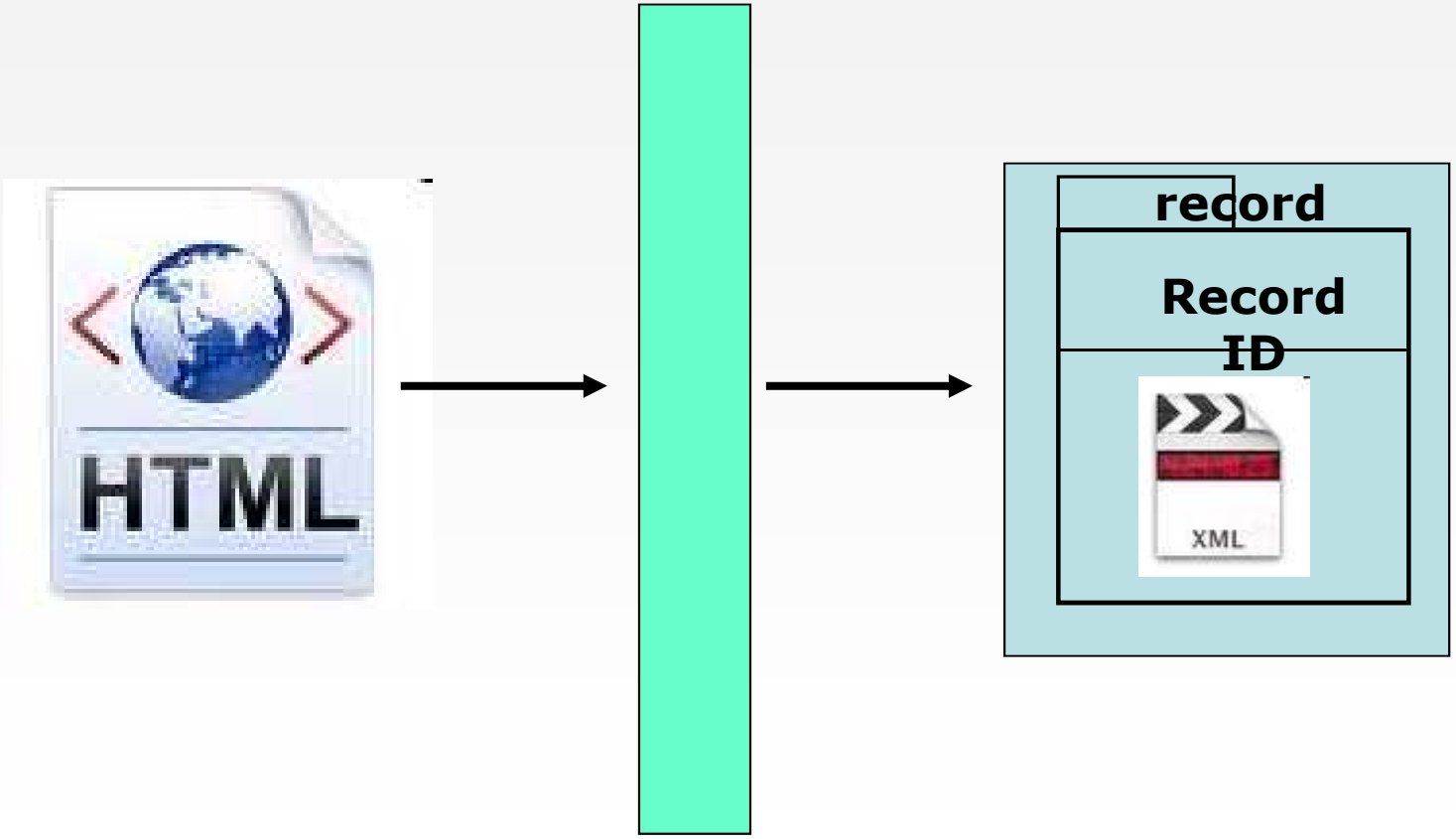
Meta tags

HTML fragment

ExLibris

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: <html> <head> **<meta name="id">**123</meta> </head> <body> </body> </html> **IdentifierXpath** = id	Yes

Additional Parameters HTML

Name	Description	Explanation
StatusXpath + StatusWhenDeleted	For example for the following file: <html> <head> **<meta name="status">**Y</meta> </head> <body> </body> </html> **IdentifierXpath** = status **StatusWhenDeleted** = Y	
ExternalResourceSourceXpath	<html> <head> **<meta name="path">** www.xxx.com/file.pdf </meta> </head> <body> </body> </html> **ExternalResourceSourceXpath**: path	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

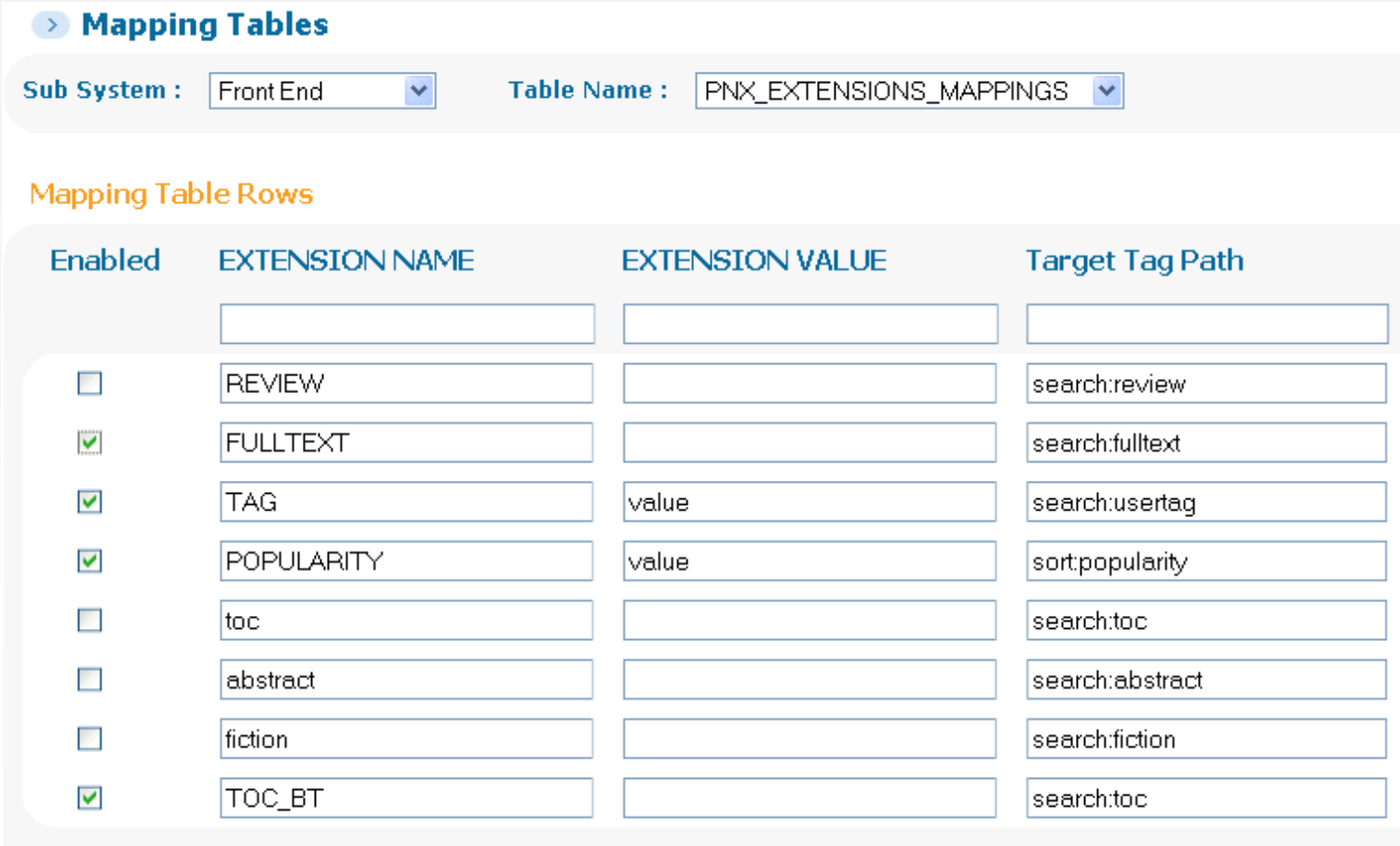
File splitters (indexing extensions)

- Must make sure 2 settings are set!!!

- No settings = **NO indexing**

File splitters indexing extensions

- Default type = FULLTEXT

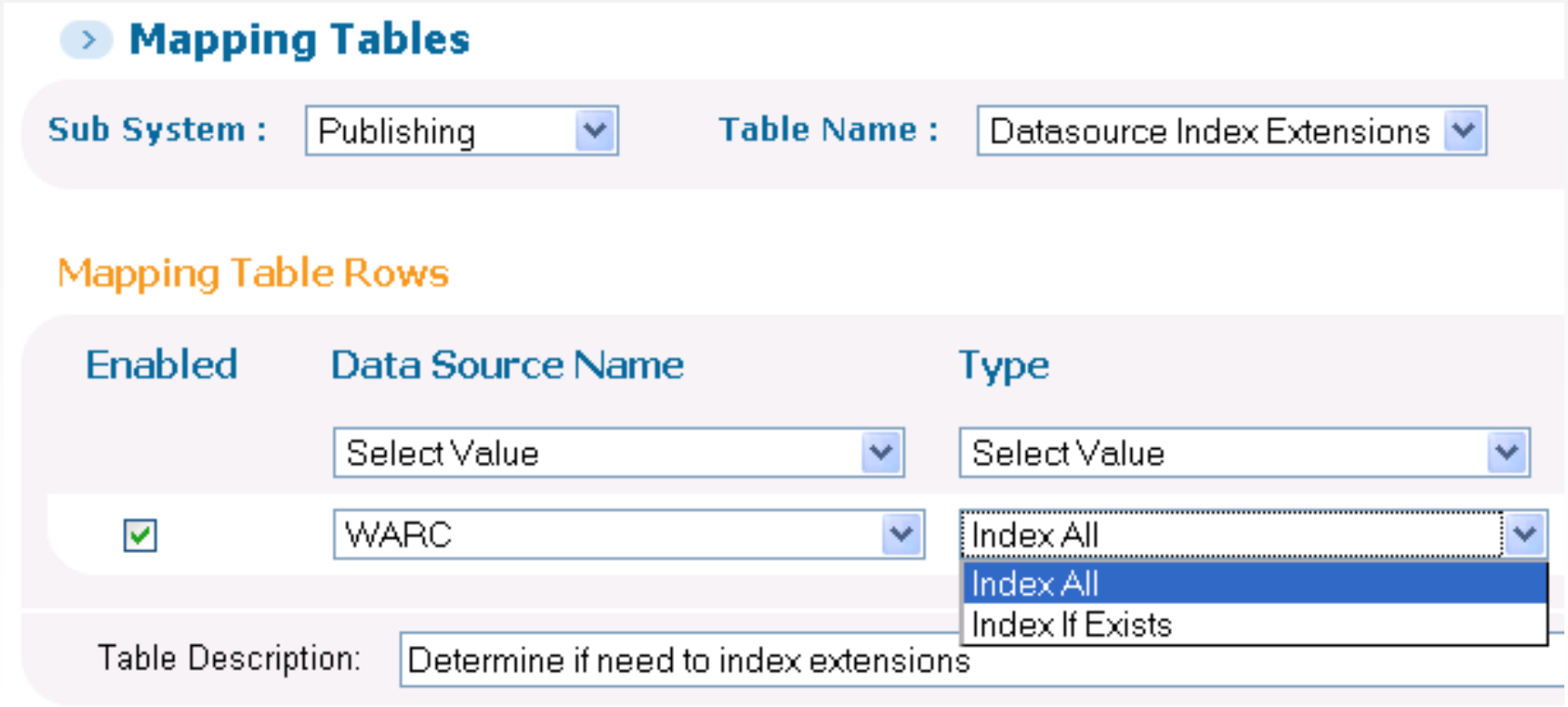


NOTE: Must enable the row!!!

ExLibris

File splitters indexing extensions

- Mark PNX with extensions

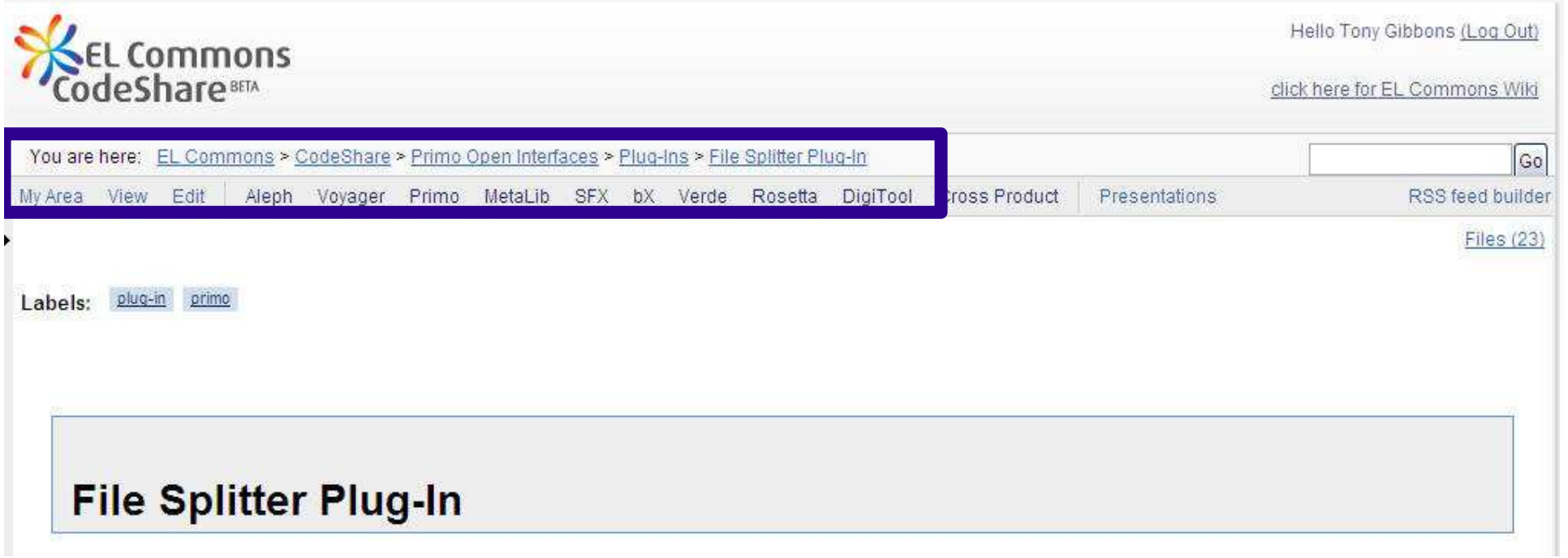


ExLibris

Documentation

EL Commons

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



ExLibris

Thank You!

Tony Gibbons, MLIS
tony.gibbons@exlibrisgroup.com

ExLibris

File splitter (Plug-in)

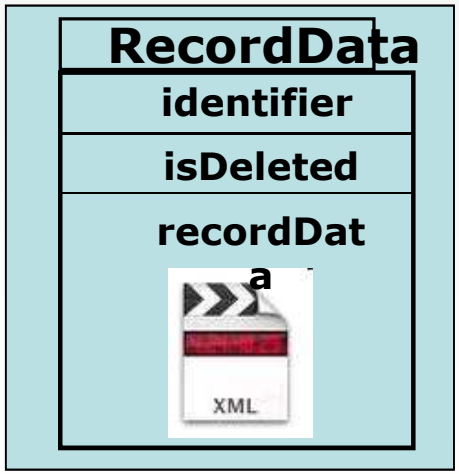
Flow:

- 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



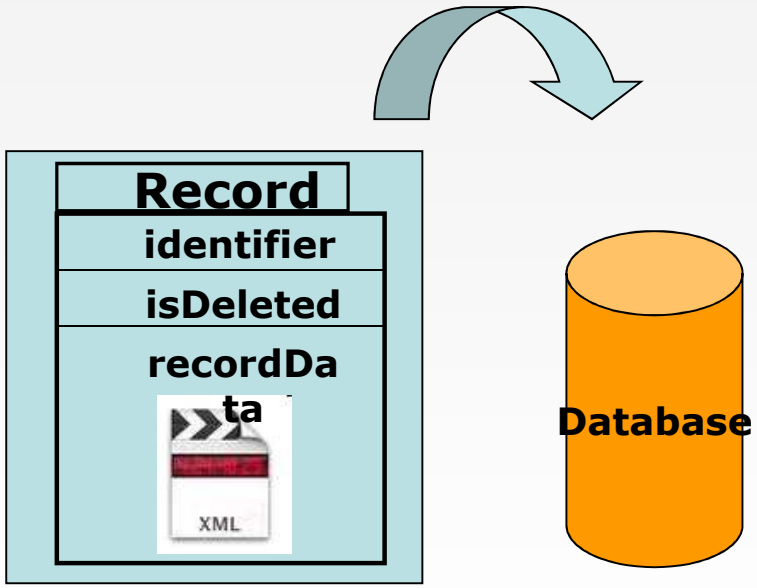
ExLibris

File splitter (Plug-in)

IRecordSaver –
Knows to save *RecordData* objects.

RecordData record =
getNextRecord();

recordSaver.save(record);



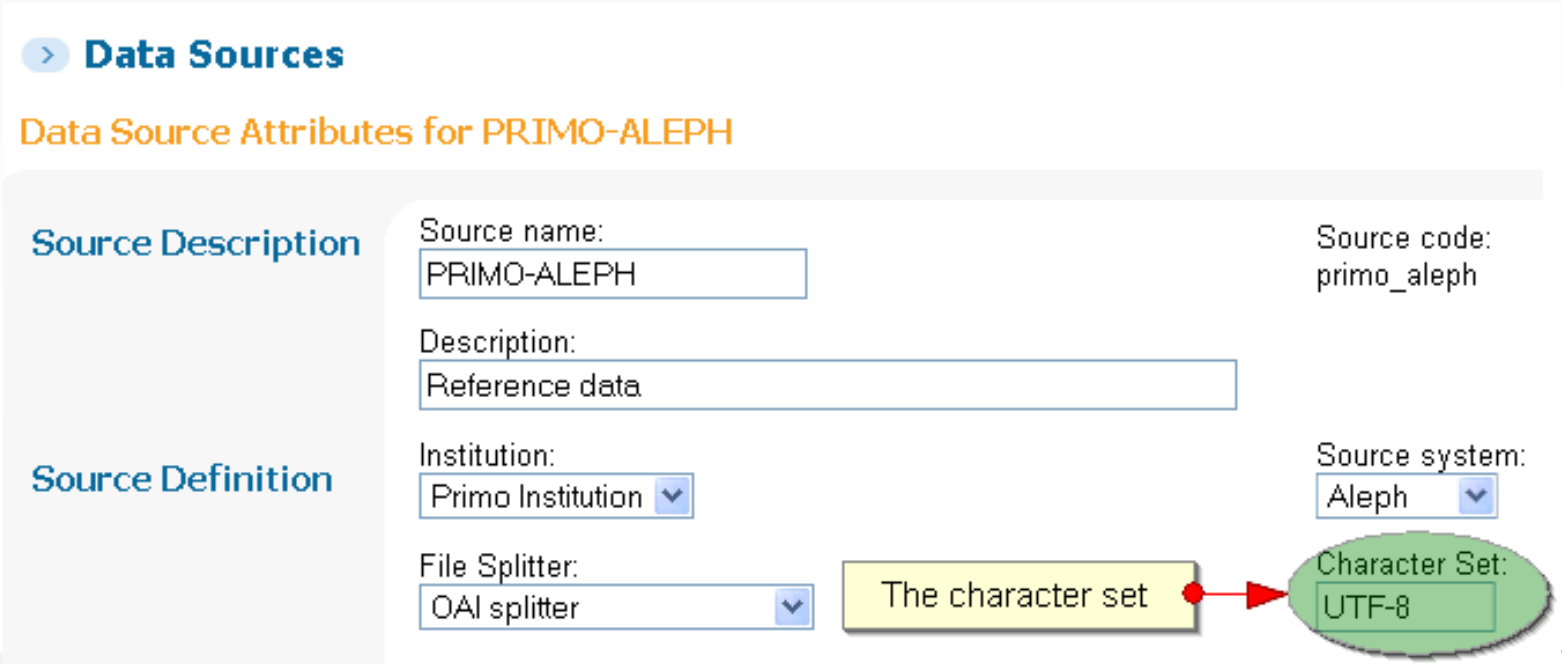
ExLibris

File splitter (Plug-in)

IFileSplitter :

#1: init()

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



File splitter (Plug-in)

IFileSplitter :

#2: parse()

1. **InputStream** –

2. **IRecordSaver** – Saves RecordData

ExLibris

File splitter (Plug-in)

IFileSplitter :

#3: doneParsing()

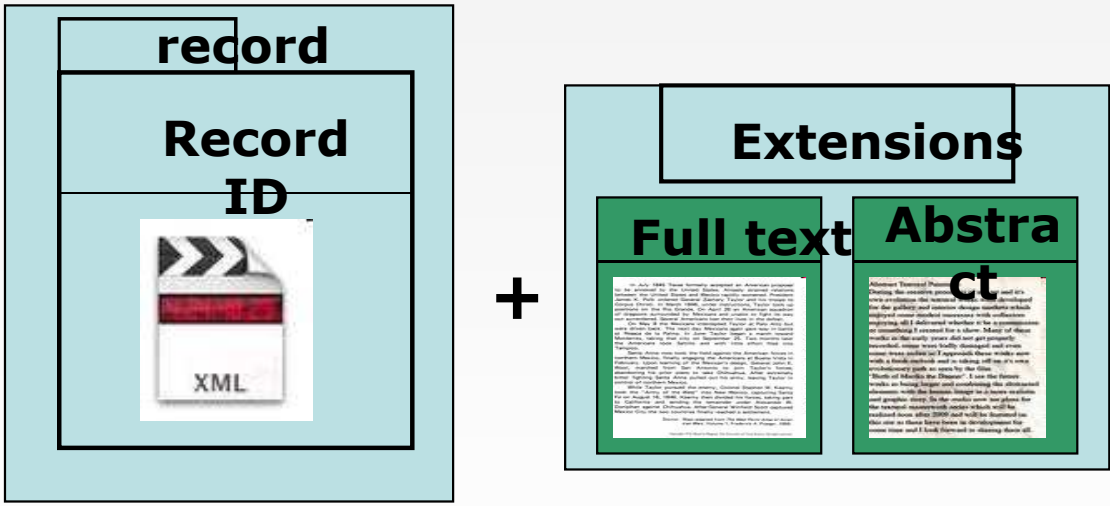
Called when no more files are left for parsing.

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);
```

ExLibris

File splitter (Plug-in)

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

#1: doBeforeSave()

1. RecordData

Boolean save = saverWrapper.doBeforeSave(rd)

if (save) {
saveRecord(rd);
}

What can be done:

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

Configure in the FS parameters

Parameter name: RecordSaverWrapper

Parameter value: A class name

ExLibris

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

ExLibris