Filter Input Records Application

# Behavior

This is a Java application that translates bibliographic record from an input format to AGRIS AP XML. This application can parse source XML files or Excel files to produce **AGRIS AP XML** files. Thus, AGRIS AP is considered the target output format. This is not because we are promoting AGRIS AP, but because we choose AGRIS AP as internal model: if we wanted to go to AGRIS RDF, we would simply need to run the Convert2RDF module to convert AGRIS AP to AGRIS RDF. The application is not generic, since a lot of customizations are needed for every input format. One could think to add also a generic module, which read a configuration file mapping input XML elements to AGRIS AP ones: anyway, no content cleaning would be possible in this situation.

The application needs to access the AGRIS SOLR INDEX to generate ARNs: ARN is a special identifier used in the AGRIS bibliographic database. This number contains information about the source where we took the information. The ARN is composed of 12 characters:

Center Code (2 chars) + Submission Year (4 chars) + Subcenter Code (1 char) + Incremental ID (5chars)

The application takes Center Code, Submission Year and Subcenter Code as command line parameters, but it needs to access the Solr index to read the last Incremental ID associated to the current source and start the generation from that number.

The application works with the following features:

* It generates 100 indented records per AGRIS AP XML output file
* ARN generation
* Duplicates removal. Duplicates are checked also by exact match on the record title. In some occasions, the user can choose if he wants to globally remove duplicates by ISSN (i.e. check if an ISSN is already in the index, skipping the record) or to accept records whose ISSNs are in the index, but associated to the same country code.
* The user can optionally specify a list of ISSNs to be filtered: only records with these ISSNs will be considered as output.
* Clean content (check ISSN syntax, add CDATA section, generation of ARN considering the last ARN in the index to avoid duplicate ARNs, etc.)

**OUTPUT: a set of AGRIS AP XML files. Each file contains at most 100 records. Output files will be placed at the same level of the source directory, in a directory called OUTPUT.**

# How to access the application

This application is a Java application whose entry point is the class:

org.fao.oekc.agris.apps.inputRecords.MainInputRecords

This class has a main method that takes as **input 7 command line parameters** the following values:

1. *The physical path of the directory recursively containing source XML files*
2. *the submission year of source records*
3. *the country code*
4. *the sub center code*
5. *a boolean flag, true for global duplicates removal, false will accept all ISSNs (if already in the index, they should be assigned to the same country code)*
6. *The format of source XML files. Supported formats are:* ***agrisap, simpledc, doaj, worldbank, ovid, mods, eldis, ajol, agricola, orbi, usamv, cirad, marcxml, bhl***
7. *(optional) The path of a text file, in which each string is separated by a new line. This file is used for different reasons. For instance, with doaj it is used to specify a list of the only ISSNs to be taken, discarding records with other ISSNs.*

The application can be executed also from Unix Bash, thanks to the script duplicatesRemoval.sh. Scripts can be generated with the build\_scripts.xml ANT script. As an example of usage:

-bash-3.2$ ./agrisInput.sh /work/agris/input/ 2012 DJ 0 false agrisap

**Note about the ARN generation:** if the parser of source files does not assign an ARN to a record, the writer class (WriteAgrisApXml.java) will generate an ARN, starting from the last ARN with a given prefix in the Solr index (retrieved by the class ArnManager.java). The incremental number is usually a 5 digits number, but the class ArnManager.java manages some exceptions: for instance, Agricola USxxxx0 needs 7 digits.

The application can run also with the bash script agrisInput.sh, tuned to work with the first six required parameters:

./agrisInput.sh /work/agris/input/agrisap/ 2012 DJ 0 true agrisap

# EXAMPLES

### agrisap

/work/agris/input/agrisap/

2012

DJ

0

true

agrisap

The AGRIS AP parser has a special behavior for the boolean flag:

* if false, it does not check for duplicates and does not generate ARNs (which should be already in tha data)
* if true, the parser checks for duplicates (based on titles) and generate new ARNs

### doaj

C:/Users/celli/Documents/workspace\_agris/agris\_input/DOAJ

2012

DJ

0

false

doaj

C:/Users/celli/Documents/workspace\_agris/agris\_input/doaj\_filter\_issn.txt

The application will accept all ISSNs present in the text file doaj\_filter\_issn.txt as list of lines, each line with an ISSN. Duplicates removal is performed on record’s title. The ARN prefix is DJ20120 and the counter starts from the last ARN in the index.

The input text file to filter ISSNs is not mandatory. So, an alternative way to run the application is:

C:/Users/celli/Documents/workspace\_agris/agris\_input/DOAJ

2015

RU

5

true

doaj

### worldbank

C:\Users\celli\Documents\workspace\_agris\agris\_input\WB\source

2012

US

4

false

worldbank

### embrapa - mods

C:\Users\celli\Documents\workspace\_agris\agris\_input\source

2012

BR

1

false

mods

### ovid

J:\AGRIS\_OVID

2012

OV

0

false

ovid

C:\Users\celli\Documents\workspace\_agris\agris\_input\cicodes.txt

In this situation, submission year, country code, and subcenter code are not relevant, since they are in the data to be converted. They are needed for the filename.

### eldis

C:\Users\celli\Documents\workspace\_agris\agris\_input\eldis\ids

2013

false

eldis

C:\Users\celli\Documents\workspace\_agris\agris\_input\eldis\eldis\_autotagger.txt

### ajol

C:\Users\celli\Documents\workspace\_agris\agris\_input\ajol\data

2013

AJ

0

false

ajol

C:\Users\celli\Documents\workspace\_agris\agris\_input\ajol\titles.txt

### agricola (ARN 7 digits counter)

C:\Users\celli\Documents\workspace\_agris\agris\_input\agricola

2013

US

0

false

agricola

### bhl

C:\Users\celli\Documents\workspace\_agris\agris\_input\data

2015

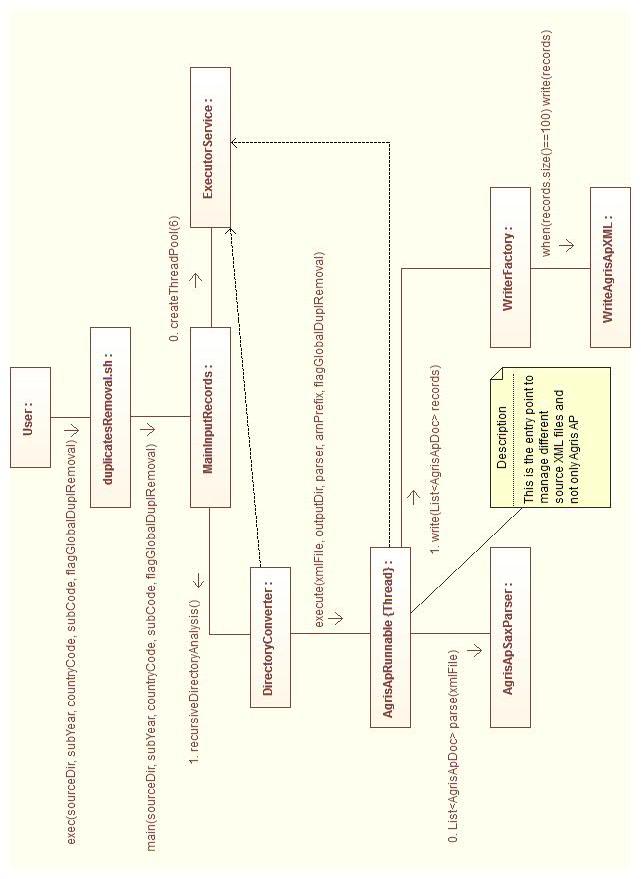
GB

4

false

bhl

# MAIN ARCHITECTURE

****

# MAPPINGS

## MODS (ORBI) – “orbi”

metadata | gs:resource ags:ARN Il code è BE 1

mods:originInfo ->

mods:dateIssued encoding="iso8601" | dcterms:dateIssued

mods:name ->

mods:namePart type="given"

mods:namePart type="family" | ags:creatorPersonal “family, given”

mods:titleInfo -> mods:title | dc:title

mods:titleInfo type="alternative || translated" ->

mods:title | dc:alternative

mods:language ->

mods:languageTerm authority="rfc3066"| dc:language scheme="dcterms:ISO639-2"

mods:originInfo><mods:publisher> | ags:publisherName

mods:classification | dc:subject {filter subjects}

mods:genre | dc:type

mods:part ->

mods:identifier type="issn" | ags:citationIdentifier scheme="ags:ISSN"

mods:part ->

mods:detail type="volume" | ags:citationNumber

mods:extent unit="pages"> | dc:extent

mods:part ->mods:detail type="series" | ags:citationTitle xml:lang=

## MODS (AGRICOLA) – “agricola”

MODS | AGRIS AP

identifier type="isbn" | dc:identifier scheme="ags:ISBN"

OR ags:citationIdentifier scheme="ags:ISSN" .

identifier type="uri" | dc:identifier scheme="dcterms:URI"

identifier type="issn" | ags:citationIdentifier scheme="ags:ISSN"

languageTerm | dc:language

name type="personal" | ags:creatorPersonal

name type="corporate"\namePart | ags:creatorCorporate

name type="conference"\namePart | ags:creatorConference

titleInfo\title | dc:title

titleInfo\subTitle | ags:titleSupplement

titleInfo type="alternative" | dcterms:alternative

titleInfo type="translated" | dcterms:alternative

place/placeTerm | ags:publisherPlace

publisher | ags:publisherName

dateIssued | dcterms:dateIssued scheme="dcterms:W3CDTF"

physicalDescription -> extent | dcterms:extent

physicalDescription ->

form authority="marcform" | dcterms:medium

note | ags:descriptionNotes

abstract | dcterms:abstract

subject/topic | dc:subject

subject authority="lcsh"\topic | ags:subjectThesaurus scheme="dcterms:LCSH" xml:lang="eng"

subject authority="lcsh"\geograp| dc:coverage\ dcterms:spatial scheme="ags:AGROVOC" xml:lang="eng"

subject authority="nal"\topic | ags:subjectThesaurus scheme="agss:NALT" xml:lang="eng"

typeOfResource | dcterms:medium

genre authority="marc" | dc:type

placeTerm type="code"

authority="marccountry" | ags:publisherPlace (concat with next tag)

placeTerm type="text" | ags:publisherPlace

dateIssued encoding="marc" | dcterms:dateIssued scheme="dcterms:W3CDTF"

location\url | dc:relation\dcterms:isPartOf scheme="dcterms:URI"

relatedItem\titleInfo\title | ags:citationTitle

relatedItem\titleInfo

type="abbreviated" | ags:citationTitle

relatedItem\part | ags:citationNumber

## MODS (BHL) – “bhl”

Similar to AGRICOLA but with less fields.

## MARC XML (VIKII) – “ marcxml”

<record> | ags:resource

<datafield tag="100"\subfield code="a" | ags:creatorPersonal

<datafield tag="700" ind1="1" <subfield code="a"> | ags:creatorPersonal

datafield tag="041"\subfield code="a" | dc:language

datafield tag="245"\subfield code="a" | dc:title English

datafield tag="260"\subfield code="b | ags:publisher name

datafield tag="260"\subfield code="c" | dcterms:dateIssued

datafield tag="773" ind1="0" <subfield code=" t"> | ags:citationTitle

datafield tag="773" ind1="0" <subfield code="g"> | The format in marc is: issue (year) space : space, volume, s. pages 21 (2012) : 4, s. 361-369

<datafield tag="773" ind1="0"<subfield code="x"> | eISSN

<datafield tag="856" ind1=" " ind2="0">

<subfield code="u"> | <dc:identifier scheme="dcterms:URI">

datafield tag="520"\ subfield code="a" | <dcterms:abstract xml:lang="eng">

datafield tag="650"\ subfield code="a"

IF code=”2” = agrovoc | agrovoc

## MODS (EMBRAPA) – “mods”

## CIRAD XML – “cirad”

## AJOL XML – “ajol”

## DOAJ XML – “doaj”

title

name

publicationDate

keyword

fullTextUrl

doi

abstract

publisher

journalTitle

issn

eissn

volume

issue

startPage

endPage

language

## ELDIS XML – “eldis”

## OVID – “ovid”

## SIMPLE DC – “simpledc”

oai\_dc:dc, oai\_qdc:qualifieddc | ags:resource ags:ARN

dc:title | dc:title

dc:creator | ags:creatorPersonal (split “ and ”)

dc:publisher | ags:publisherName

dc:contributor | ags:creatorCorporate

dc:date, dcterms:issued | (no buffer) dcterms:dateIssued

dc:language | (no buffer) dc:language

dc:subject | dc:subject (split “/”,”,”,”;”)

dc:identifier | dc:identifier dcterms:URI

dc:relation | dc:relation\dcterms:isPartOf scheme="dcterms:URI"

dc:type | (no buffer) dc:type

dc:source | dc:source

dc:rights | dc:rights

dc:format | (no buffer) dc:format/dcterms:medium

dc:coverage | (no buffer) dc:coverage

dc:description | dcterms:abstract

Customizations:

* abstract: GB (DFID), AV (AVANO)
* date: GB (DFID), AV (AVANO)
* identifier: GB (DFID), AV (AVANO)
* relation: GB (DFID), AV (AVANO), US (AGECON)
* format: GB (DFID)
* language: US (AGECON)
* coverage: AV (AVANO)
* source: AV (AVANO)
* excluded: AV FAO

## AGRIS AP – “agrisap”

If the parameter *globalDuplicatesRemoval* is true, the parser checks for duplicates (based on titles) and generate new ARNS. If false, it does not check for duplicates and does not generate ARNs (which should be already in the data)

## WORLD BANK EXCEL – “worldbank”

## ROMANIA USAMV XML – “usamv”

**Default language title, abstract: en**

<Article> | ags:resource ags:ARN

<PublisherName> | ags:publisherName

<JournalTitle> | ags:citationTitle xml:lang=en

<Issn> | ags:citationIdentifier scheme="ags:ISSN"

<Year>2013</Year> | dcterms:dateIssued

<Volume>70</Volume> | ags:citationNumber

<FirstPage>267</FirstPage>

<LastPage>276</LastPage> | ags:extent

<Author>

<FirstName>Cecilia</FirstName>

<LastName>Alexandri</LastName> | ags:creatorPersonal

<ArticleTitle> | dc:title xml:lang=en

<Language> | dc:language scheme="dcterms:ISO639-2"

<Abstract> | <dcterms:abstract xml:lang="en">