**Authors:**

Lucile Grand (Ministère de la Culture et de la Communication, France)
Yoann Moranville (Ministère de la Culture et de la Communication, France)

# Table of content

# Introduction

The OAI console harvester is a standalone tool that you can easily install on any computer. It has been developed to enable you to:
- test the OAI access to your repository in the Archives Portal Europe, once existing
- harvest the whole contents of your repository in the APE. While the automatic harvesting function of the Content checker limits the harvest to the first ten records of any set of your repository and the dashboard only harvests at night, the OAI console tool allows you to work completely independent of the limitations of the Dashboard.

The harvester is also used by Europeana to harvest our APE data

The console works just like the harvester included in the back-office of the portal, with some more options. Basically, it follows the OAI protocol that is based on an exchange of questions and answers between the harvester and the repository.
Requirement: in order to use the standalone tool, you need to have Java 6 ore a more recent version installed on your computer. You can verify which version is installed on your computer by launching the command line (Windows) or terminal (UNIX) and typing *java -version*. You also need an internet connection to actually harvest data from a repository.

# 1. install and launch the tool

The tool is distributed as a ZIP file, named OAIHarvester-*2.0.3*-package, where the numbers indicate the version of the tool.
Updated version of the tool is available for registered users at https://redmine.archivesportaleurope.net/projects/apex/wiki/Servers_and_releases while the most recent version "2.0.10" can be downloaded at http://dpt.archivesportaleurope.net/OAIHarvester-2.0.10-package.zip

Please do the following once you have the ZIP file:
• Uncompress the ZIP-file with winzip, 7-zip, unzip or a similar tool on a location of your choice to create the executable and auxiliary files (2 folders (*conf* and *lib*) and 5 files).

• To launch the program please perform the following steps according to your operating system (OS):
- Using Windows: Double-click on the *oaiharvester.bat* file,
- Using UNIX: Using your terminal launch the *oaiharvester.sh* file (i.e. Type *shoaiharvester .sh* – make sure that the file has at least the execution rights, *chmod +x oaiharvester .sh*),
- Using OS X: Double-click on the *OAIHarvester.jar* file[1].

• The program should be now running (see figure below). If it is not the case, please refer to the third chapter of this manual.

---

[1] Double-clicking the *OAIHarvester.jar* file can also be used as an alternative to the *oaiharvester.bat* file with Windows.

OAI console harvester launched: you are asked to provide the address of the repository
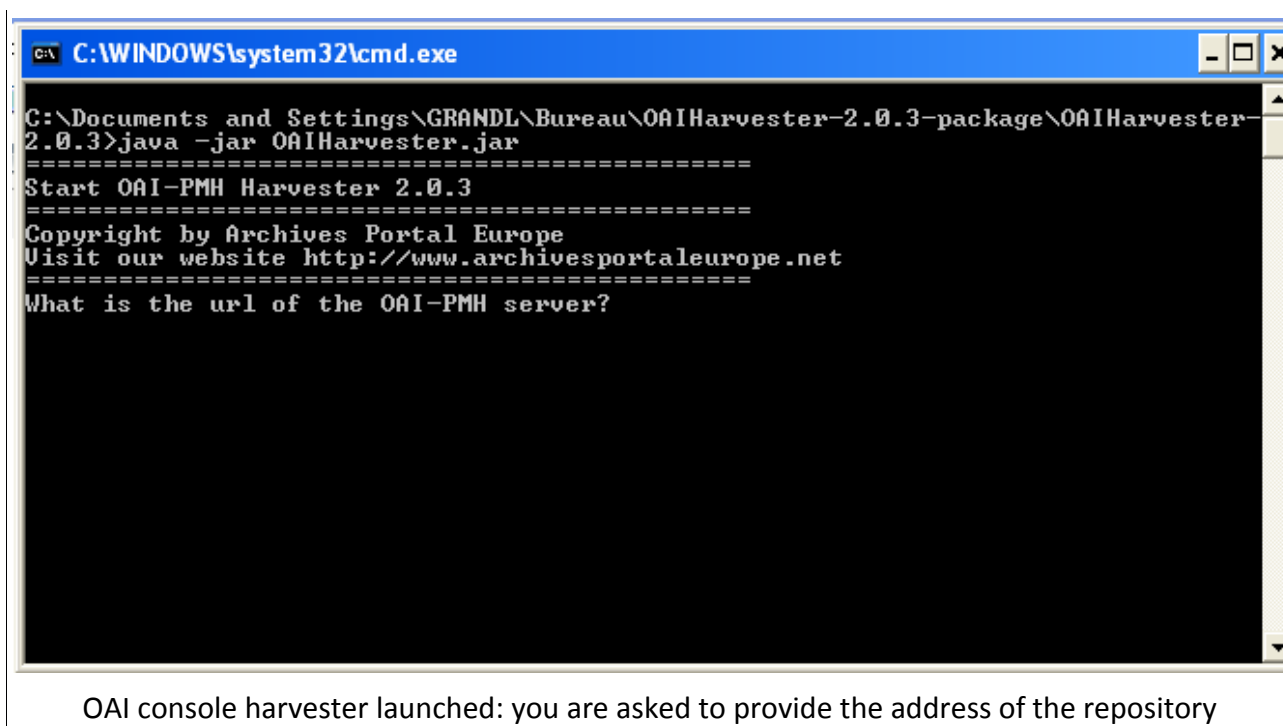
• When launched, the program creates two additional folders, *data* and *logs* containing the harvested data as well as the log files where the actions processed will be registered (useful when errors occur).



List of files when the tool has been launched

In the *logs* folder, the *harvester* file indicates in detail how the harvests have been proceeded and allow to check everything if needed

## 2. Set up the harvest

To set up the harvest, you just have to follow the instructions displayed on the screen.
How does the tool function? The harvester sends the requests to the repository by using the normal OAI-PMH syntax (beginning with the first, the verb Identify) and proposes the choice between the different possibilities offered by the repository as soon as it receives the answers.

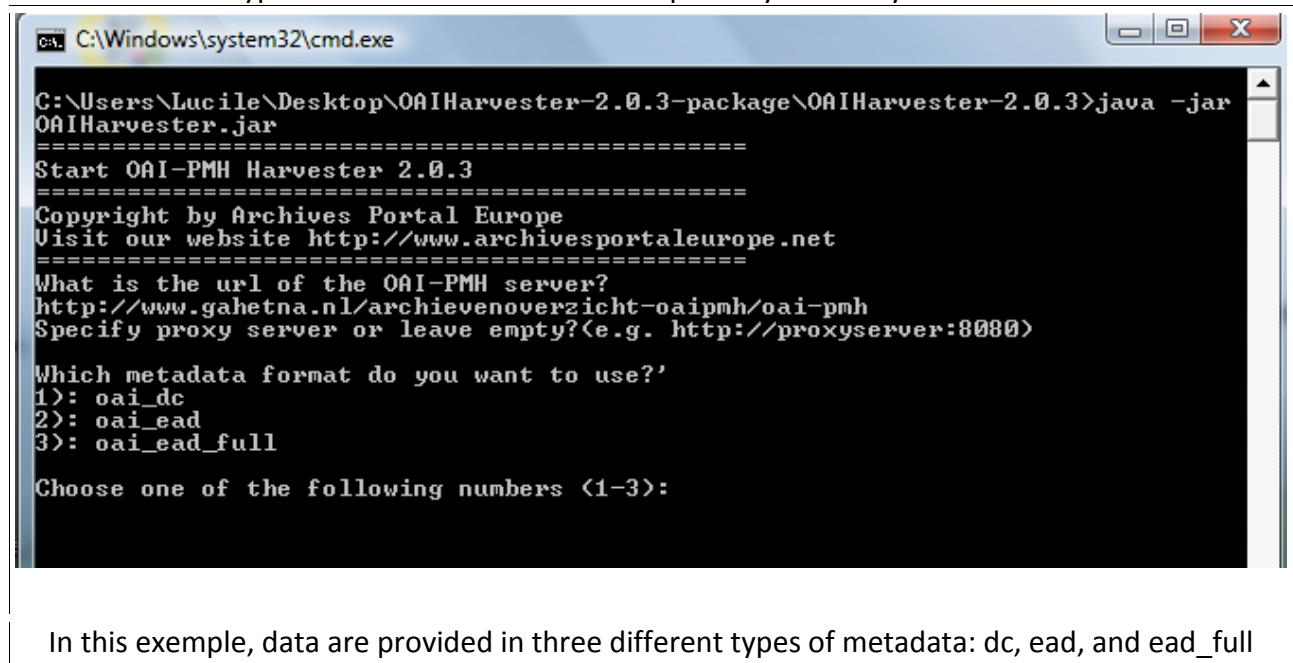### 2.1 indicate the address of your repository

The address must include the protocol: http:// or https://
for instance: http://www.archives-nationales.culture.gouv.fr/oai/

The console harvester then asks you to indicate a proxy (it can be mandatory in some environments, for security reasons) or to leave empty. To leave empty, just press the enter key.

The harvester begins its « dialog » with the repository by sending the request verbs and providing the according answers: list of metadata, list of sets, etc.

## 2.2 select the type of metadata that you want to harvest

The tool lists the types of metadata found in the repository and aks you to select one of them.



In this exemple, data are provided in three different types of metadata: dc, ead, and ead_full

## 2.3 select the set that you want to harvest

The tool lists the sets found in the repository and gives them an arbitrary number to allow you to choose one.
Please note that you can harvest only one set after another; if you want to harvest everything you have operate set by set.

```
C:\Windows\system32\cmd.exe
What is the url of the OAI-PMH server?
http://archives.cantal.fr/oai_pmh.cgi
Specify proxy server or leave empty?(e.g. http://proxyserver:8080)

Which metadata format do you want to use?'
1): oai_dc
2): ead
3): olac

Choose one of the following numbers (1-3): 2
Which set do you want to use?'
1): documents_figures (Documents figurús)
2): sons_et_videos (Sons et vidúos)
3): notaires (Notaires)
4): test_naoned (test Naoned)
5): sources_imprimees (Sources imprimúes)
6): dictionnaire_topographique (Dictionnaire topographique)
7): etat_civil_ (Etat civil)
8): cadastre_napoleonien (Cadastre napolúonien)
9): tables_et_repertoires_:_registres_des_hypotheques (Tables et rúpertoires : r
egistres des HypothÞques)
10): autres_archives_numerisees (Autres archives numúrisúes)
11): cadre_de_classement (Cadre de classement)
12): 45_fi (45 Fi)

Choose one of the following numbers (1-12):
```

In this exemple, you have 12 thematic sets



```
C:\Windows\system32\cmd.exe
Copyright by Archives Portal Europe
Visit our website http://www.archivesportaleurope.net
=====================================
What is the url of the OAI-PMH server?
http://www.gahetna.nl/archievenoverzicht-oaipmh/oai-pmh
Specify proxy server or leave empty?(e.g. http://proxyserver:8080)

Which metadata format do you want to use?'
1): oai_dc
2): oai_ead
3): oai_ead_full

Choose one of the following numbers (1-3): 2
Which set do you want to use?'
1): naa1 (Nationaal Archief û Archives category 1)
2): naa2 (Nationaal Archief û Archives category 2)
3): naa3 (Nationaal Archief û Archives category 3)
4): naa4 (Nationaal Archief û Archives category 4)

Choose one of the following numbers (1-4):
```
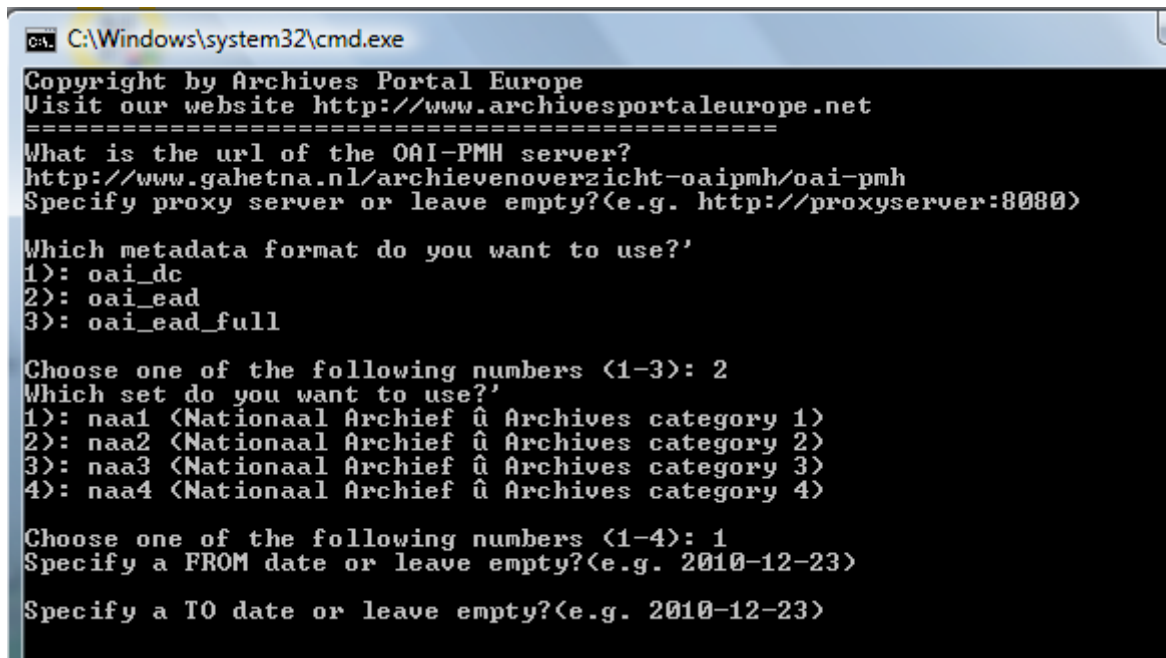
In this exemple, you have 4 "numeric" sets

## 2.4 select the intervall of dates

This is not mandatory, and useful only to make a differential harvest.



In this exemple, the FROM and TO date have been left empty by pressing the enter key

## 2.5 select the harvest method

The usual harvesting method is by ListRecords; but in some case, you might have to use the ListIdentifiers/GetRecord combination verbs

## 2.6 select the type of records that you want to save

You can choose to save either the metadata record (so the simple original file), either the full OAI response (so the original file included in the OAI "wrapper" ; see figure below).
Once harvested the records are saved in the data folder, either in the oai sub-folder, either in the ead-subfolder.

Choose the type of files to save



Exemple of a full OAI response: you see that the ead file is enclosed in the oai response

## 2.7 launch the harvest

The OAI console sums-up your choices so that you can check your demand before actually launch the harvest.



```
==================================================
Summary of OAI-PMH Harvester parameters
==================================================
Url of the OAI-PMH server:              http://www.gahetna.nl/archievenoverzicht
-oaipmh/oai-pmh
Metadata format:                        oai_ead
Set:                                    naa1
Harvest method:                         Harvest by verb ListRecords
Store method:                           Save full OAI-PMH responses
Location of the files to be stored:     C:\Users\Lucile\Desktop\OAIHarvester-2.0
.3-package\OAIHarvester-2.0.3\data\www.gahetna.nl\oai_ead\naa1
==================================================
Do you want to proceed?
1): Yes
2): No
```

Summary of the requests

The harvest begins: the files are retrieved by groups of two (each answer contains two files). Each line in the figure below represents one file harvested: (1,1) indicates that this is the first record of the first response, (1,2) the second record of the first response.



```
C:\Windows\system32\cmd.exe

Choose one of the following numbers (1-2): 1
==================================================
Summary of OAI-PMH Harvester parameters
==================================================
Url of the OAI-PMH server:              http://recherche.archives.manche.fr/oai_
pmh.cgi
Metadata format:                        ead
Set:                                    son
Harvest method:                         Harvest by verb ListIdentifiers/GetRecor
d (fail safe)
Store method:                           Save only the metadata record (e.g. EAD,
  EDM or DC files)
Location of the files to be stored:     C:\Users\Lucile\Desktop\OAIHarvester-2.0
.3-package\OAIHarvester-2.0.3\data\recherche.archives.manche.fr\ead\son
==================================================
Do you want to proceed?
1): Yes
2): No

Choose one of the following numbers (1-2): 1
(1,1): LI(u): 'FRAD050_000016' (2015-01-16)
(1,2): LI(u): 'FRAD050_000065' (2013-12-10)
(2,3): LI(u): 'FRAD050_000066' (2013-12-10)
(2,4): LI(u): 'FRAD050_000067' (2013-12-12)
```

In this exemple, four files have already been harvested in two responses

## 2.8 retrieve the files

The harvested files are placed in the *data* folder, hierarchically ordered by repository, type of files (oai or simple file) and set.

You can then process them in the Data preparation tool for further tests or check if they have been properly handled when placed in your repository (for instance if some information has to be

added automatically to your files such as the full urls to link back to your data etc.)
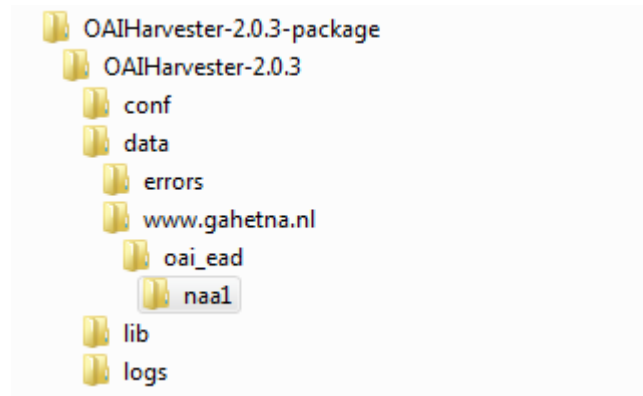


data folder including the possible errors and the files

# 3. troubleshooting, possible errors and difficulties

If problems were to be found, please send us an email via
http://www.archivesportaleurope.net/contact and select "Technical issues" in the proposed subjects.
If the tool contains bugs or you suspect it does, please do the following before sending an email:
• Go to the output directory of the tool and retrieve the file named *errors.log*,
• Copy and paste the content of that file report in the email with some explanations on the error that occurred.

The console provides an error report when needed – displayed in the command line window and – or as a text file in the errors folder.

```
harvester - Bloc-notes
Fichier  Edition  Format  Affichage  ?
09:57:00,910 INFO  ================================================
09:57:00,910 INFO  ================================================
09:57:00,925 INFO  Start OAI-PMH Harvester 2.0.3
09:57:00,925 INFO  Start OAI-PMH Harvester 2.0.3
09:57:00,925 INFO  ================================================
09:57:00,925 INFO  ================================================
09:57:00,925 INFO  Copyright by Archives Portal Europe
09:57:00,925 INFO  Copyright by Archives Portal Europe
09:57:00,925 INFO  Visit our website http://www.archivesportaleurope.net
09:57:00,925 INFO  Visit our website http://www.archivesportaleurope.net
09:57:00,925 INFO  ================================================
09:57:00,925 INFO  ================================================
10:09:37,386 INFO  ================================================
10:09:37,386 INFO  ================================================
10:09:37,386 INFO  Summary of OAI-PMH Harvester parameters
10:09:37,386 INFO  Summary of OAI-PMH Harvester parameters
10:09:37,386 INFO  ================================================
10:09:37,386 INFO  ================================================
10:09:37,386 INFO  Url of the OAI-PMH server:        http://www.gahetna.nl/archievenoverzicht-oaipmh/oai-pmh
10:09:37,386 INFO  Url of the OAI-PMH server:        http://www.gahetna.nl/archievenoverzicht-oaipmh/oai-pmh
10:09:37,386 INFO  Metadata format:                  oai_ead
10:09:37,386 INFO  Metadata format:                  oai_ead
10:09:37,386 INFO  Set:                              naa1
10:09:37,386 INFO  Set:                              naa1
10:09:37,386 INFO  Harvest method:                        Harvest by verb ListRecords
10:09:37,386 INFO  Harvest method:                        Harvest by verb ListRecords
10:09:37,402 INFO  Store method:                          Save full OAI-PMH responses
10:09:37,402 INFO  Store method:                          Save full OAI-PMH responses
10:09:37,402 INFO  Location of the files to be stored:  C:\Users\Lucile\Desktop\OAIHarvester-2.0.3-package\OAIHarvester-2.0.3\data\www.gahetna.
10:09:37,402 INFO  Location of the files to be stored:  C:\Users\Lucile\Desktop\OAIHarvester-2.0.3-package\OAIHarvester-2.0.3\data\www.gahetna.
10:09:37,402 INFO  ================================================
10:09:37,402 INFO  ================================================
10:11:38,288 INFO  0 records harvested successfully with no errors.
10:11:38,288 INFO  0 records harvested successfully with no errors.
10:11:38,288 INFO  ================================================
10:11:38,288 INFO  ================================================
10:11:38,288 INFO  Elapsed time: 0 hour(s) 0 minute(s) 1 second(s)
10:11:38,288 INFO  Elapsed time: 0 hour(s) 0 minute(s) 1 second(s)
10:11:38,288 INFO  ================================================
10:11:38,288 INFO  ================================================
10:11:38,288 INFO  OAI-PMH Harvester 2.0.3 finished
10:11:38,288 INFO  OAI-PMH Harvester 2.0.3 finished
10:11:38,288 INFO  ================================================
10:11:38,288 INFO  ================================================
10:11:38,303 INFO  Copyright by Archives Portal Europe
10:11:38,303 INFO  Copyright by Archives Portal Europe
10:11:38,303 INFO  Visit our website http://www.archivesportaleurope.net
10:11:38,303 INFO  Visit our website http://www.archivesportaleurope.net
10:11:38,303 INFO  ================================================
10:11:38,303 INFO  ================================================
10:11:52,983 INFO  ================================================
10:11:52,983 INFO  ================================================
10:11:52,998 INFO  Start OAI-PMH Harvester 2.0.3
10:11:52,998 INFO  Start OAI-PMH Harvester 2.0.3
```

Exemple of *Harvester* file