

ISTI Open Access - Support #26104

Reconsider the harvesting pipeline to take into account IRIS contents

Nov 27, 2023 11:37 AM - Leonardo Candela

Status:	Closed	Start date:	May 30, 2024
Priority:	Normal	Due date:	
Assignee:	Michele Artini	% Done:	100%
Target version:			
VREName:			
Description			
The endpoint to use is <a href="https://iris.cnr.it/oai/openaire4">https://iris.cnr.it/oai/openaire4</a> with metadataPrefix=oai_openaire			
For ISTI-CNR the set is ou_ou294			
Subtasks:			
Task # 27587: IRIS oaire:resourceType management			Closed

History

#1 - Apr 19, 2024 05:56 PM - Leonardo Candela

- Subject changed from Test harvesting from the forthcoming IRIS-based CNR Repository to Test harvesting from IRIS CNR Repository
- Priority changed from Normal to Urgent

According to the last information we get we should try to harvest from

<https://iris.cnr.it/oai/openaire4>

and use the oai\_openaire metadata format.

The ISTI set is ou\_ou294

There are a couple of known limitations:

- the information we will collect per author is limited (in the near future we will have an ID for CNR Authors);
- there is no information about previously used People IDs (in the near future they will give us a mapping file);

The files, if open access are exposed with their own URLs. Let's think about the impact (e.g. we can get rid of the local FTP, we have to reconsider Matomo)

#2 - May 08, 2024 08:56 PM - Leonardo Candela

- Priority changed from Urgent to Immediate

We managed to get a file containing the mapping between the "old" People IDs and the new IRIS IDs ... this file is available at <https://data.d4science.net/4F4b>

Concerning author information, it seems that we should count of snippets like this

```
<datacite:creator>
  <datacite:creatorName>Candela L</datacite:creatorName>
  <datacite:nameIdentifier schemeURI="https://orcid.org" nameIdentifierScheme="ORCID">0000-0002-7279-2727
</datacite:nameIdentifier>
  <affiliation affiliationIdentifierScheme="ROR" affiliationIdentifier="https://ror.org/04zaypm56">
National Research Council</affiliation>
</datacite:creator>
<datacite:creator>
  <datacite:creatorName>Hedges M</datacite:creatorName>
</datacite:creator>
```

NB. all CNR authors have the CNR affiliation rather than the specific institute affiliation.

#3 - May 09, 2024 08:07 AM - Michele Artini

- Status changed from New to In Progress

NB. all CNR authors have the CNR affiliation rather than the specific institute affiliation.

It 's OK. Also People did not have the affiliations to the institute. The ISTI OpenPortal has always showed all the CNR authors related to the publications in the ISTI oai set.

The new mapping rules will be integrated as soon as possible in the BETA OpenPortal.

**#4 - May 29, 2024 05:27 PM - Leonardo Candela**

- % Done changed from 0 to 50

**#5 - May 29, 2024 05:36 PM - Leonardo Candela**

- Subject changed from *Test harvesting from IRIS CNR Repository* to *Reconsider the harvesting pipeline to take into account IRIS contents*

- Description updated

- % Done changed from 50 to 40

Changed the description of the ticket to reflect the current settings.

For the time being, the results are available by <https://openportal.beta.isti.cnr.it>

**#6 - May 29, 2024 05:36 PM - Leonardo Candela**

- Description updated

**#7 - May 30, 2024 12:51 PM - Leonardo Candela**

I kindly as the following:

- we should enable the *daily harvesting*, so that we see how things are evolving;
- we should reconsider the *venue* where the publication is appearing, we no longer have a pre-cooked string.
  - for *journal article* we should use `citationTitle`
  - for the rest of products I need to have a look

**#8 - May 30, 2024 01:21 PM - Michele Artini**

I scheduled the aggregation wf to run each day at 22:30.

**#9 - Dec 23, 2024 10:03 AM - Leonardo Candela**

- Status changed from *In Progress* to *Closed*

The new pipeline is there