invenio-oaiharvester Documentation

Release 0.1.0

CERN

August 19, 2015

Contents

Invenio module for OAI-PMH metadata harvesting between repositories.

- Free software: GPLv2 license
- Documentation: https://invenio-oaiharvester.readthedocs.org.

This is an experimental development preview release.

CHAPTER 1

Features

This module allows you to easily harvest OAI-PMH repositories, thanks to the Sickle module, and feed the output into your ingestion workflows, or simply to files. You can configure your OAI-PMH sources via a web-interface and run or schedule immediate harvesting jobs via command-line or regularly via Celery beat.

CHAPTER 2

Harvesting is simple

```
inveniomanage oaiharvester get -u http://export.arxiv.org/oai2 -i oai:arXiv.org:1507.07286 > my_recor
```

This will harvest the repository for a specific record and print the records to stdout - which in this case will save it to a file called `my_record.xml`.

If you want to have your harvested records saved in a directory automatically, its easy:

```
inveniomanage oaiharvester get -u http://export.arxiv.org/oai2 -i oai:arXiv.org:1507.07286 -o dir
```

Note the output `-o` parameter that specifies how to output the harvested records. The three options are:

- Sent to a workflow (E.g. *-o workflow*)
- Saved files in a folder (E.g. *-o dir*)
- Printed to stdout (default)

CHAPTER 3

Harvesting with workflows

```
inveniomanage oaiharvester get -u http://export.arxiv.org/oai2 -i oai:arXiv.org:1507.07286 -o workflo
```

When you send an harvested record to a workflow you can process the harvested files however you'd like and then even upload it automatically into your own repository.

This module already provides some

CHAPTER 4

Managing OAI-PMH sources

If you want to store configuration for an OAI repository, you can use the administration interface available via the admin panel. This is useful if you regularly need to query a server.

Here you can add information about the server URL, metadataPrefix to use etc. This information is also available when scheduling and running tasks:

```
inveniomanage oaiharvester get -n somerepo -i oai:example.org:1234
```

Here we are using the *-n, –name* parameter to specify which stored OAI-PMH source to query, by name.

CHAPTER 5

API

If you need to schedule or run harvests via Python, you can use our API:

```python
from invenio_oaiharvester.api import get_records
for rec in get_records(identifiers=["oai:arXiv.org:1207.7214"],
                       url="http://export.arxiv.org/oai2"):
    print rec.raw
```