Metadata curation: hands-on session

CMDI and Metadata Curation Task Forces

CLARIN Centre & Developers meeting
4-5 June 2018
Utrecht, The Netherlands

Prerequisites

- Java 1.8
  - https://java.com/en/download/
- Internet connection

Menu

1. View your records in the VLO
2. View your harvest (and its log)
3. Get your records
    1. From the tarball
    2. Harvest them
4. Curation module
    1. Look at the website
    2. Run locally
5. CMDI best practices
    1. Check your profiles
    2. Check your records
6. Structural queries
    1. Load your records/validation reports into BaseX
    2. Some useful XQueries
7. Inspect the mapping
8. VLO
9. Fixing problems, but where?
10. What's missing?

View your records in the VLO

- Filter the records based on your endpoint:
  - `_oaiEndpointURI`:
    - https://vlo.clarin.eu/search?q=_oaiEndpointURI:https://clarin-pl.eu/oai/request
  - Endpoints? centres.clarin.eu/oai_pmh
- Filter the records based on a profile:
  - `_componentProfile`:
    - https://vlo.clarin.eu/search?q=_componentProfile:LINDAT_CLARIN
    - *Note*: use the profile name instead of its ID!

View your harvest (and its log)

- Not in production yet, but local preview
  - will replace https://vlo.clarin.eu/data/

- Paged lists
- Filter on endpoints and/or records
- See the log of a harvest

Get your records

1. From the tarball
   1. https://vlo.clarin.eu/data/resultsets/
   2. `tar xjf clarin.tar.bz2 results/cmdi/DANS_CMDI_Provider`
   3. *Note*: just clicking the tarball might freeze your Mac!
2. Harvest them
   1. https://github.com/clarin-eric/oai-harvest-manager/releases
   2. Edit providers section of resources/config-test.xml
   3. `run-harvester.sh workdir=`pwd` resources/config-test.xml`

Curation module

1. Look at the website
    1. https://clarin.oeaw.ac.at/curate/
2. Run locally
    1. https://github.com/clarin-eric/clarin-curation-module
    2. curation.jar (goo.gl/Cx4h3N)
    3. Create your own specific copy of config.properties
    4. `java -jar curation.jar -config config.properties -c -path results/cmdi/ARCHE`

CMDI best practices

1. https://www.clarin.eu/content/cmdi-best-practice-guide
2. Schematron rules (schematron.com)
   1. https://github.com/TheLanguageArchive/SchemAnon/releases
   2. Also supported by oXygen or other XML editors
   3. Also easy to define your own rules
3. Check your profiles
   1. Identify the profiles you're using
      1. https://github.com/clarin-eric/FindProfiles/releases
      2. `java -jar findProfiles.jar -e=xml clarin/results/cmdi/The_Language_Archive/`
   2. `wget -O profile.xml https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1505397653795/xml && java -jar SchemAnon.jar https://raw.githubusercontent.com/clarin-eric/cmdi-toolkit/develop/src/main/resources/toolkit/sch/cmd-component-best-practices.sch profile.xml`

CMDI best practices

4. Check the records
    1. `java -jar SchemAnon.jar https://raw.githubusercontent.com/clarin-eric/cmdi-toolkit/develop/src/main/resources/toolkit/sch/cmd-record-best-practices.sch clarin/results/cmdi/The_Language_Archive/xml`
    2. Note: use the `-s` option to save the SVRL report
5. Validate the records
    1. https://github.com/clarin-eric/cmdi-instance-validator/releases
    2. `cmdi-validator results/cmdi/IMS_Repository/`
    3. Note: use the `-s` option to use another Schematron file

Structural queries

1. Load your records/validation reports into BaseX
   1. basex.org or `brew install basex`
   2. Create a new database and import your records/reports
2. XQuery (w3.org/XML/Query)

```
declare namespace  cmd="http://www.clarin.eu/cmd/1";
declare namespace  svrl="http://purl.oclc.org/dsdl/svrl";
```

…

- goo.gl/CEZtTm

Notes

1. You can use the namespace wildcard (`*:element`) to deal with (many) profile specific namespaces
2. You can use `base-uri()` to get the file name of a matching record
3. BaseX has useful modules, but also FunctX (xqueryfunctions.com)
4. A problem that occurs often might be a candidate for a Schematron rule

Inspect the mapping

1. Identify the profiles you're using
    1. https://github.com/clarin-eric/FindProfiles/releases
    2. `java -jar findProfiles.jar -e=xml clarin/results/cmdi/The_Language_Archive/`
2. Inspect the mapping
    1. https://github.com/clarin-eric/VLO-mapping
    2. https://cmdi.clarin.eu/mapping/

VLO

1. Curation VLO (to be updated)
    1. https://vlo.minerva.arz.oeaw.ac.at/vlo
2. Request an import in the beta VLO
    1. vlo@clarin.eu
3. Do a local VLO import
    1. https://gitlab.com/CLARIN-ERIC/compose_vlo#run-the-importer-to-ingest-cmdi-metadata-into-the-vlo
4. Run the importer on one record
    1. https://github.com/clarin-eric/VLO/blob/master/vlo-importer/src/main/java/eu/clarin/cmdi/vlo/importer/MetadataMapper.java
        - `CLASSPATH="vlo-importer-4.2-SNAPSHOT-importer.jar" java eu.clarin.cmdi.vlo.importer.MetadataMapper -c VloConfig.xml -r test.xml`

Fixing problems, but where?

- Your records
  - Typos in your records
  - Inconsistencies in your records
    - Consider adopting a common (CLARIN/CLAVAS) vocabulary
  - Facet mapping problems
    - Can you fix them in your profile(s)?
    - Or provide feedback to the Metadata Curation TF (cmdi@clarin.eu)
  - Value mapping problems
    - Provide feedback to the Metadata Curation TF (cmdi@clarin.eu)
- Others records
  - report them via the VLO feedback button

What's missing?

- OAI Viewer
  - history
  - include a local curation run
  - general log
  - mail to technical contact when number of harvested records drop
- VLO importer
  - report showing the mappings applied
- VLO
  - `_componentProfileURI`
    - profile name might not be unique!
  - Center facet
    - filter to all records from one center, possible multiple endpoints
  - show original value
- More?

Questions

- Metadata Curation Taskforces
  - tf-curation@lists.clarin.eu
- CMDI Taskforce
  - cmdi@clarin.eu
- CMDI first aid kit
  - clarin.eu/sites/default/files/CMDI-first-aid-kit.pdf
- Menzo Windhouwer
  - menzo.windhouwer@di.huc.knaw.nl