VLO development meeting

The Netherlands, 15 - 19 January 2018

This document serves as an **agenda** and for **note keeping**. For the schedule, locations and other details, see the meeting page.

Pre-meeting (Monday afternoon)

For early arrivals :)

When: Monday afternoon
Where: Nijmegen, office
Who: Matej & Twan; perhaps also Menzo and/or Dieter

- Time for ad-hoc discussions and preparing the agenda

VLO/curation/metadata quality meeting (Tuesday)

Focus on curation and quality side of the VLO and CMDI, als looking forward to the curation workshop in Vienna later this month.

When: Tuesday ~11:15[1] - 17:00
Where: Utrecht
Who: Dieter (part of the time), Matej, Twan, Wolfgang, (Menzo?) - also lunch(?) with Franciska and Maria of CLARIN

1. Curation
    a. Curation workflow
    See https://github.com/clarin-eric/VLO-mapping
        i. Motivation
        ii. Conceptual recap
        iii. Run-through with examples
    b. MD Quality Dashboard
    (See "Variability of the Facet Values in the VLO - a Case for Metadata Curation", section 5.2, p. 36)
    https://trac.clarin.eu/wiki/VLO/CMDI%20data%20workflow%20framework
    => deploy new OAI-harvest viewer [Menzo]

[1] Or when Wolfgang arrives

=> add Data Provider / Data Endpoint facet in VLO

Actions on the datasets:

i. ~~(Re-)harvesting of the data set~~

ii. ~~Disable indexing~~

iii. ~~Delete the data set~~

iv. Show the curation report

- data quality report (see Section 5.3) (download them as XML, PDF, etc.)
- Show the error messages (download them as PDF etc.)
- Show the schema/profile (with the link to Component Registry, CLAVAS, and CCR)

v. Show the metadata records => in VLO, or as raw files in the harvester

vi.

vii. ~~Send an email to the data provider (e.g. data quality report)~~

-> should be triggered by some automated action

Or even we may want to restrain from sending lot of automated emails

viii. Run curation module as part of harvesting

2. Value mapping/normalisation
   a. Value Mapping I - value facet mapping proposal
      (Note: technical aspects should be discussed in next day's meeting (VLO development with Thomas present)
   b. (What are the others doing?
      - d-net@CNR@PARTHENOS, (basic info)

      - MINT@ATHENA@Europeana)

[Lunch]
Franciska will join, (probably?) in the early afternoon. Agenda (order) tentative.

3. Curation c'td
   a. Curation module
      (See CLARIN+ D2.1)
      i. General - State and development
      ii. Code sharing with VLO?
      iii. Features?
      iv. Use by Assessment Committee
      v. Integration with other services such as CLAVAS

      => weighting of the scoring
4. Vocabularies
   a. CLAVAS Current state
      OpenSKOS has user base and user community
      Editor messy code (with its own
      Technically: OpenSKOS2 (skos:collection, handles, but not editor) just use API
      Is just triple store and server with clear API

    b. Vocabs.dariah.eu on SKOSMOS (run by ACDH-OEAW)
       => share evaluation spreadsheet
       Also used in Parthenos
    c. Vocabulary Editors
       Not so needed in CLAVAS and CCR
       Currently not a priority (based on experience) rather use more agile approach
       THEMAS
    d. API
       New REST-API being designed - on github [Menzo]
       Common API for SKOSMOS and OpenSKOS?
       => compare the APIs new OpenSKOS
       Or take ELDA on top SPARQL (Menzo tried with SISSvoc)
    e. Which ones (licenses/availability/license categories; resource type)
       Look into vocabularies by Parthenos for inspiration (CERIF)
       CLARINO-licences overview
       2017-06-28 - Curation Meeting on licenses
       LicenseAvailibilityMap.gsheet
       Current license - availability map
       Europeana: rightstatements.org
       Alternative approach suggestion: Can I access it/Can I use it?
5. Upcoming curation workshop in Vienna
   => weighting of the scoring
   => investigate curation score of BBAW
6. URL checking
    a. Number of URLs:
       6.1mio resource refs (1mio records have 2 resource refs, most of remainder have 1)
       1.1mio landing pages
    b. Balancing of the request to prevent accidental DDOSsing and being blacklisted

We may have some additional time e.g. for hands-on work.

Misc resources

- Europeana Data Quality Committee has some interesting presentations illustrating their issues and approach

Dinner and/or drinks in Nijmegen (to meet Thomas)?

Action points

This week

- [Menzo, Twan, Wolfgang] Implement compilation module prototype for curation workflow (CLAVAS + CSV ===Stylesheets===> uniform map
- [Twan] Share curation workflow drawing source
- [Matej, Twan, ..?] prepare vocabularies/mappings for the curation workflow
- [Matej] share SKOS evaluation spreadsheet
- [Menzo] integrate curation tool into the harvester

Soon

- [Menzo] Deploy harvest viewer
- [Menzo] export mapping from directory name to endpoint, center, national project
    - [Twan] Add field to VLO for OAI endpoint ('data provider')
- [?] Design and/or prototype for link checker (based on curation module logic?)
- Design for metadata quality dashboard

Later

- [Curation module] Field-weighting in curation module
- [Curation module] investigate curation score of BBAW
- [Menzo, ?] Adapt harvester viewer to fully fledged MD quality dashboard

---

VLO development meeting day 1 (Wednesday)

General VLO development meeting to discuss process, progress and planning

When: Wednesday 9:00 - 17:00
Where: Sky lounge, Nijmegen (E.20.26, Erasmus Building)
Who: Dieter, Matej, Menzo, Wolfgang, Thomas, Twan

1. VLO evaluation/vision/brainstorm
    a. Aims
        i. Original aim vs current aim of the VLO
        LREC 2012 paper*: In the era of the digital data deluge, a researcher needs efficient ways to navigate to the language resources that really matter, whatever the selection criterion is. A plethora of resource inventories and catalogues has been proposed to address this need. However, almost all of them are based on a single metadata scheme, forcing the resource providers to trade off accuracy in favor of compatibility.* (Also see LREC 2010 paper)

        LREC 2010 poster*: The Virtual Language Observatory provides multiple views on metadata for linguistic data and software tools. In analogy with the astronomical virtual observatories, it tries to give a consistent online overview of the data and tools that are available at a large variety of resource centres worldwide. Metadata must be open to allow everyone to find useful resources and tools. VLO is simply one portal that is harvesting all available metadata and giving the credits to the providers.*
        ii. Changes in (intended) use since inception?
        iii. Current fitness in terms of original/current aim
    b. Current use cases
        i. Within CLARIN
        ii. Outside CLARIN

c. Rethinking the VLO
[How would we reimplement the VLO if time were not an issue? Functional and technical aspects...]
Inspiration: http://www.osaarchivum.org/digital-repository; https://www.europeana.eu/portal/en/search?q=

2. VLO development process
   a. Distribution of work and responsibilities
   b. Release cycle
   c. Development workflow
      i. Use of GitHub (issues, milestones, pull requests, code review)
   d. Infrastructure (dev, testing, production environments - sufficient?)
   e. Use of docker
   See https://gitlab.com/CLARIN-ERIC/compose_vlo for info and pointers
      i. In deployments
      See https://docs.google.com/spreadsheets/d/1mVk8J1wMVfNUwCYgaKpC0ZqyvkMhVJTQahyg0Nn6xKI/edit?usp=sharinghttps://github.com/clarin-eric/VLO/blob/master/DEPLOYMENT.md
      ii. In development
      See https://github.com/clarin-eric/VLO/blob/master/DEVELOPMENT.md
3. VLO (technical) issues
   Discussions related to the VLO importer, web-app, etc
   a. Reported issues
   Review of issues reported by colleagues from Slovenia
      i. Document: https://drive.google.com/file/d/1NbLXmC1_EHqF9dimZfEn2L62qz7HSmrv/view?usp=sharing
      ii. Issue overview https://docs.google.com/spreadsheets/d/1mVk8J1wMVfNUwCYgaKpC0ZqyvkMhVJTQahyg0Nn6xKI/edit?usp=sharing
   b. Value mapping II
   https://github.com/clarin-eric/VLO/issues/93 (needs a better/more comprehensive issue description)

   c. VloConfig.xml schema
   See https://github.com/clarin-eric/VLO/issues/124
   d. Possible support of alternative “value selection widgets” (e.g. for numeric/temporal values, hierarchical values, geographical selection etc.)
      i. => primary resource type filter(s)
   e. images for records/collections/organisation
      i. https://github.com/clarin-eric/VLO/issues/54 is related
   f. Icons for resource types
4. EOSC-hub
5. IEEE Hackathon

6. Vocab/CSV -> value mapping prototype
7. Brainstorm 2.0
    a. Rethinking the VLO
    [How would we reimplement the VLO if time were not an issue? Functional and technical aspects...]

Action points

- Investigate the creation of landing pages for specific languages (manually curated + links to VLO)
- Add VLO and FCS search to drupal search widget on www.clarin.eu (ask Hendrik?)
- Icons per collection to visualize?
- Issue for supporting Java 9
- Darja & Jakob report
    - Look into weird querying behaviour “german OR dutch” vs “(german OR dutch)”
        - Also look at lower case ‘or’
    - Find info on “copy/paste problem” in old solr logs (e.g. ‘Talk of Norway’)
    - Looking for corpora/collections flooded with individual item, aka forest-trees problem aka granularity problem, possible approaches:
        - Similar record folding
        - Try to explicitly distinguish collection(corpus)/item records
    - Find use cases for value mapping where context would allow for better mapping
        - For example ‘type’ field in olac
    - Olac-linguistic-type map to resource type rather than genre
        - But also look into other cases where genre mapping might be justified
- Common library/tool for link checker (see todos of 16.1)
- EDM-CMDI conversion:
    - resolve library of congress subject heading codes to preflabel
    - Also getty for resource type e.g. http://vocab.getty.edu/aat/300026656
- See if it is possible to suggest discriminative additional search terms wrt query results
- Resource type facet: see if OTA values (154 distinct values) can be dealt with through concept mapping (blacklist?) and/or value mapping
- Resourcetype: collection vs. corpus discussion
    - Make them synonyms in solr-configuration
    - Corpus narrower term of collection -> hierarchical facet
    - “Corpus” false positive Corpus Vitraeum
- Think about result sorting feature and/or discuss with Jakob
- Problem statement for value mapping

Dinner options

- Bella Italia
- **Credible**
- Indian Way
- De Hemel

- Stoom
- ...

VLO development meeting day 2 (Thursday)

Further discussions related to the VLO importer, web-app, etc

When: Thursday 9:00 - 11:00
Where: Nijmegen (Sky lounge and/or CLARIN offices)
Who: Dieter, Thomas (Morning), Twan, Wolfgang (Menzo virtually)

14:00 Vidconf with Menzo about value mapping/curation prototype

---

Knowledge transfer and hands-on 2 (Friday)
Time reserved for knowledge transfer and/or working on concrete issues together.

- Curation module
  - VLO code/logic sharing
- VLO
  - XSD for VloConfig.xml
  - Look at other open issues in GitHub
- 10:45 VLO dev planning w/ Dieter
  - VLO 4.4
  - VLO 4.5
- Any other knowledge transfer?

- 12:00 Lunch?

- Value mapping solution
  - Adapt CFM to new definition structure
  - 14:00 Video conference with Menzo

Action points

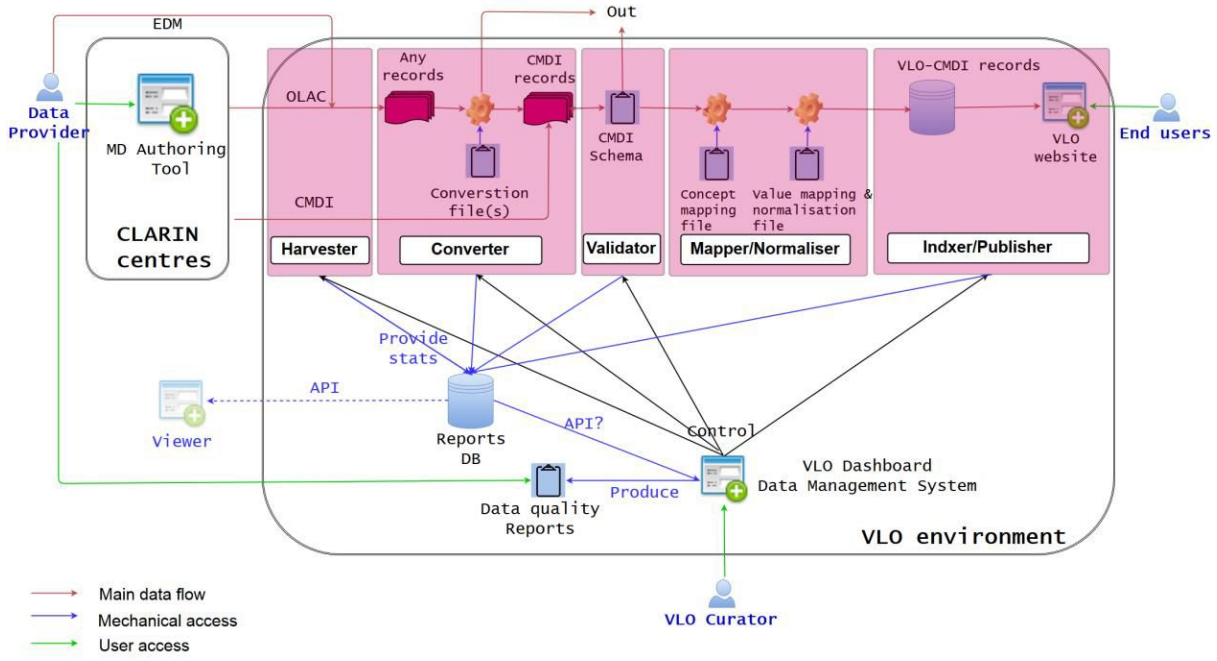
- This week
  - Try to replace logic to get external VloConfig in curation module
    - Ideal solution would be to 'harvest' it from running production VLO
  - Move issue https://github.com/clarin-eric/VLO/issues/46 to VLO-mapping
  - Create github issue for integrating url checker (results) into the VLO (providing input to 'can I access it' -see #...)
- Soon
- Later
  - Curation module: use VLO importer as a library to share the mapping and processing logic rather than have it duplicated

When: Friday 9:00 - 16:00
Where: Nijmegen (CLARIN offices)
Who: Twan, Wolfgang (Menzo?, Dieter?)

Topics: see above

Afternoon: Wolfgang meeting with Daan

King, Sugimoto, Ostojic, Durco (2015): Variability of the Facet Values in the VLO – a Case for Metadata Curation
http://www.ep.liu.se/ecp/123/003/ecp15123003.pdf



https://trac.clarin.eu/attachment/wiki/VLO/Meeting/2017-01-16/VLO%20dynamic%20mapping%20workflow.png