The Virtual Language Observatory

Dieter Van Uytvanck

CLARIN-FR tutorial, Paris

Overview

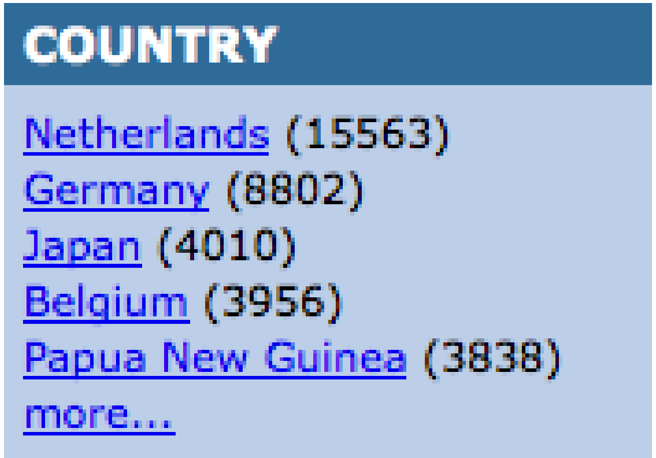
- VLO?
- What is behind it?
- How do I get my data in there?
- Demo

VLO?

- Virtual Language Observatory
- http://www.clarin.eu/vlo/
- Several parts:
  - Google Earth overlay
  - Facet browser:
    - Data
    - Tools (to be integrated into the data browser)
  - Catalogue
  - LRT inventory

Facets?

- A simple way to narrow down the search space, step by step
- Purpose: quickly navigating through a huge amount of metadata
- **Not** the tool to answer research questions



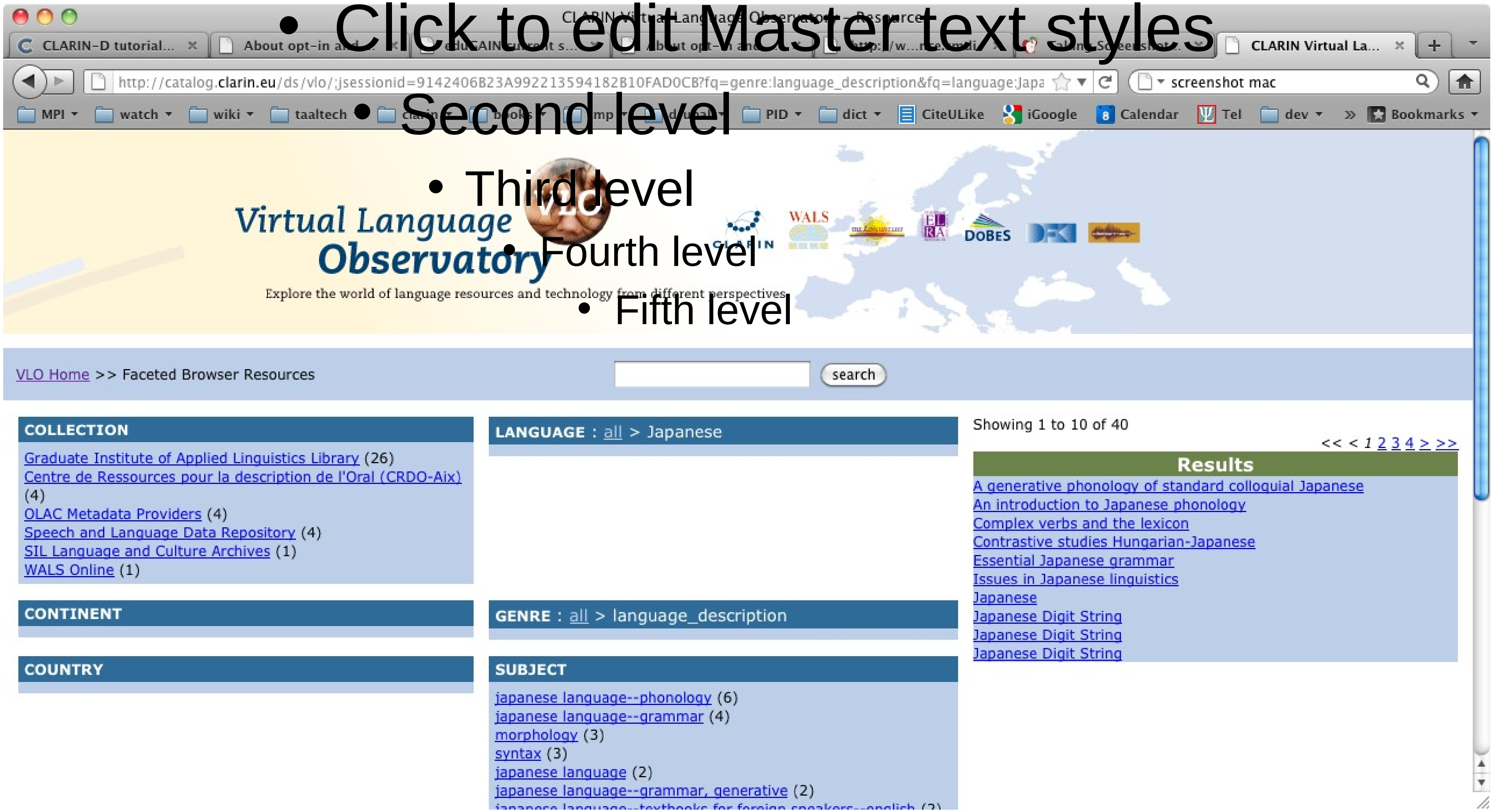
COUNTRY

Netherlands (15563)
Germany (8802)
Japan (4010)
Belgium (3956)
Papua New Guinea (3838)
more...

VLO Faceted Browser (1)

CLARIN-D

- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level


CLARIN Virtual Language Observatory - Resources
CLARIN-D tutorial...
About opt-in and...
CLARIN Virtual La...
http://catalog.clarin.eu/ds/vlo/;jsessionid=9142406B23A992213594182B10FAD0CB?fq=genre:language_description&fq=language:Japa
screenshot mac
MPI
watch
wiki
taaltech
PID
dict
CiteULike
iGoogle
Calendar
Tel
dev
Bookmarks
Virtual Language Observatory
Explore the world of language resources and technology from different perspectives
WALS
DOBES

VLO Home >> Faceted Browser Resources

search

COLLECTION

Graduate Institute of Applied Linguistics Library (26)
Centre de Ressources pour la description de l'Oral (CRDO-Aix) (4)
OLAC Metadata Providers (4)
Speech and Language Data Repository (4)
SIL Language and Culture Archives (1)
WALS Online (1)

CONTINENT

COUNTRY

LANGUAGE : all > Japanese

GENRE : all > language_description

SUBJECT

japanese language--phonology (6)
japanese language--grammar (4)
morphology (3)
syntax (3)
japanese language (2)
japanese language--grammar, generative (2)
japanese language--textbooks for foreign speakers--english (2)

Showing 1 to 10 of 40

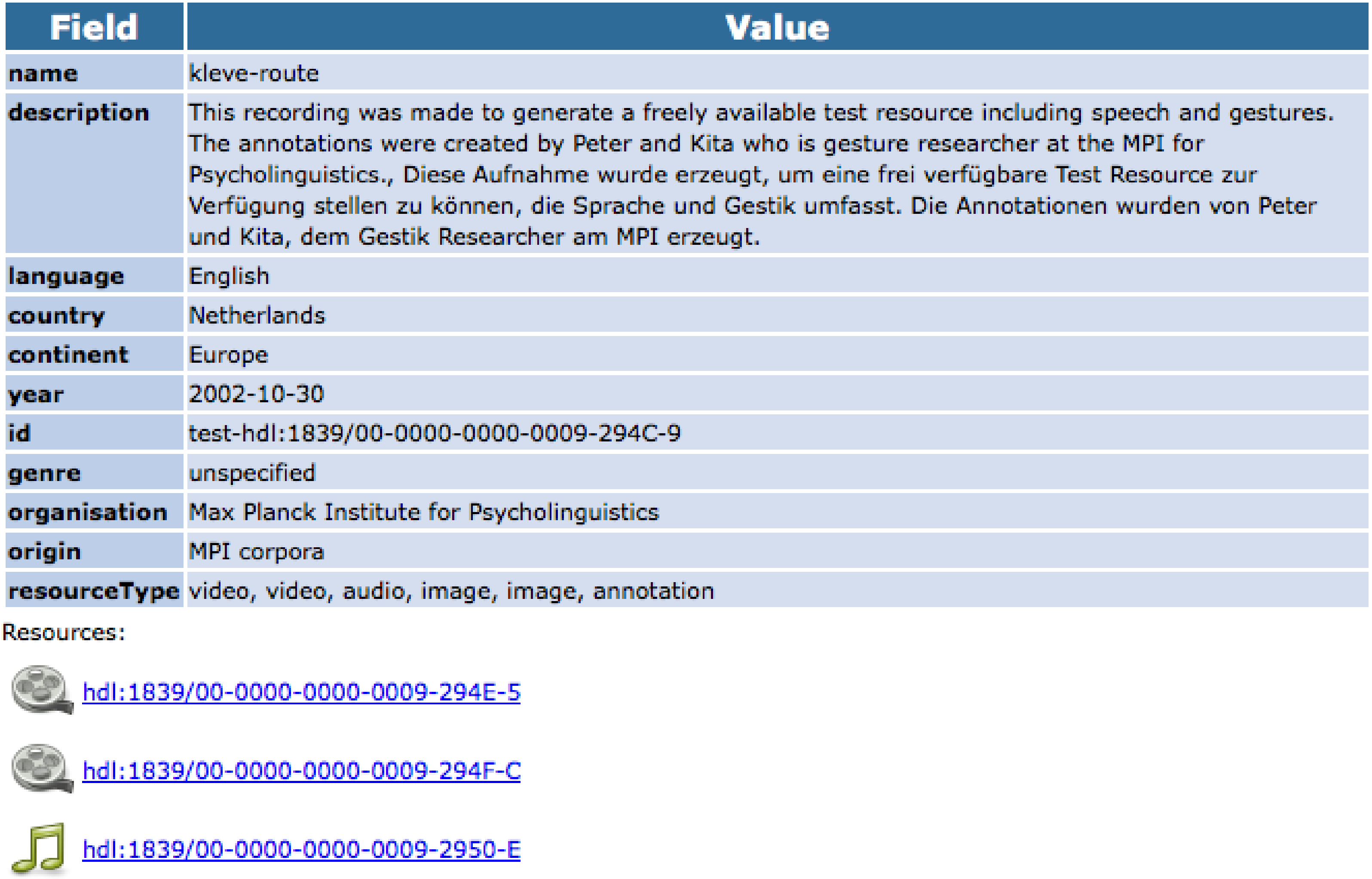
<< < 1 2 3 4 > >>

Results

A generative phonology of standard colloquial Japanese
An introduction to Japanese phonology
Complex verbs and the lexicon
Contrastive studies Hungarian-Japanese
Essential Japanese grammar
Issues in Japanese linguistics
Japanese
Japanese Digit String
Japanese Digit String
Japanese Digit String

http://catalog.clarin.eu/ds/vlo

VLO Faceted Browser (2)



Field	Value
name	kleve-route
description	This recording was made to generate a freely available test resource including speech and gestures. The annotations were created by Peter and Kita who is gesture researcher at the MPI for Psycholinguistics., Diese Aufnahme wurde erzeugt, um eine frei verfügbare Test Resource zur Verfügung stellen zu können, die Sprache und Gestik umfasst. Die Annotationen wurden von Peter und Kita, dem Gestik Researcher am MPI erzeugt.
language	English
country	Netherlands
continent	Europe
year	2002-10-30
id	test-hdl:1839/00-0000-0000-0009-294C-9
genre	unspecified
organisation	Max Planck Institute for Psycholinguistics
origin	MPI corpora
resourceType	video, video, audio, image, image, annotation

Resources:

🎞 hdl:1839/00-0000-0000-0009-294E-5

🎞 hdl:1839/00-0000-0000-0009-294F-C

♫ hdl:1839/00-0000-0000-0009-2950-E

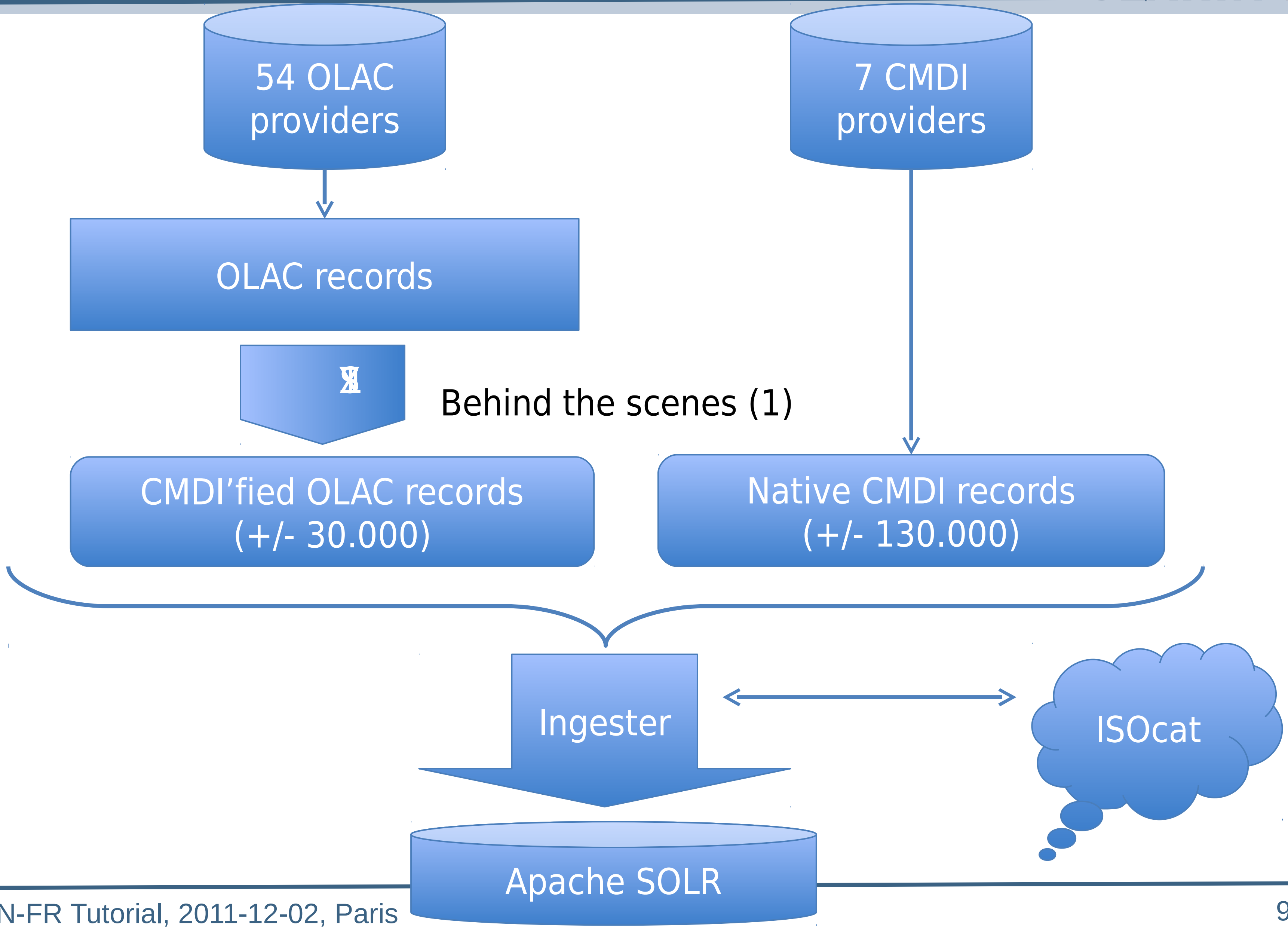
VLO Faceted Browser (3)

- Data analyzed is CMDI format
- Data sources
  - CMDI files harvested from CLARIN centres
  - CMDI'fied OLAC records (from CLARIN centres and others)
  - CMDI'fied IMDI
  - CMDI'fied LRT inventory records
- Stability issue solved and performance improved
- You can get to resources directly from search results

Behind the scenes (1)

- SOLR + lucene
- Tomcat web application
- For the parsing of the CMDI’s: VTD-XML
  - Faster than SAX-parser
  - Still full XPath access
  - Memory-efficient (1.3x~1.5x the size of an XML document)

Behind the scenes (2)



54 OLAC providers

OLAC records

Behind the scenes (1)

CMDI’fied OLAC records (+/- 30.000)

7 CMDI providers

Native CMDI records (+/- 130.000)

Ingester

ISOcat

Apache SOLR

VLO and ISOcat: natural allies (1)

- The import of metadata files used to be hard coded
- Now we look at the ISOcat links in the XSDs as generated from the CMDI profiles
- Fallback to hard-coded XPath in case no ISOcat link is found

VLO and ISOcat: natural allies (2)

- Import configuration example:

```
<facetConcept name="name" allowMultipleValues="false">

<concept>http://www.isocat.org/datcat/DC-2544</concept> <concept>
http://www.isocat.org/datcat/DC-2545</concept> <concept>http://purl.org/dc/terms/title
</concept>

<!-- no concept in lrt schema -->
<pattern>
/c:CMD/c:Components/c:LrtInventoryResource/c:LrtCommon/c:ResourceName/text()
</pattern>

</facetConcept>
```

How do I get my data in there?

- Provide it as CMDI over OAI-PMH
- Provide it as OLAC over OAI-PMH
- Provide it as IMDI over OAI-PMH
- Enter it into the LRT inventory:
  - www.clarin.eu/inventory

Orde r

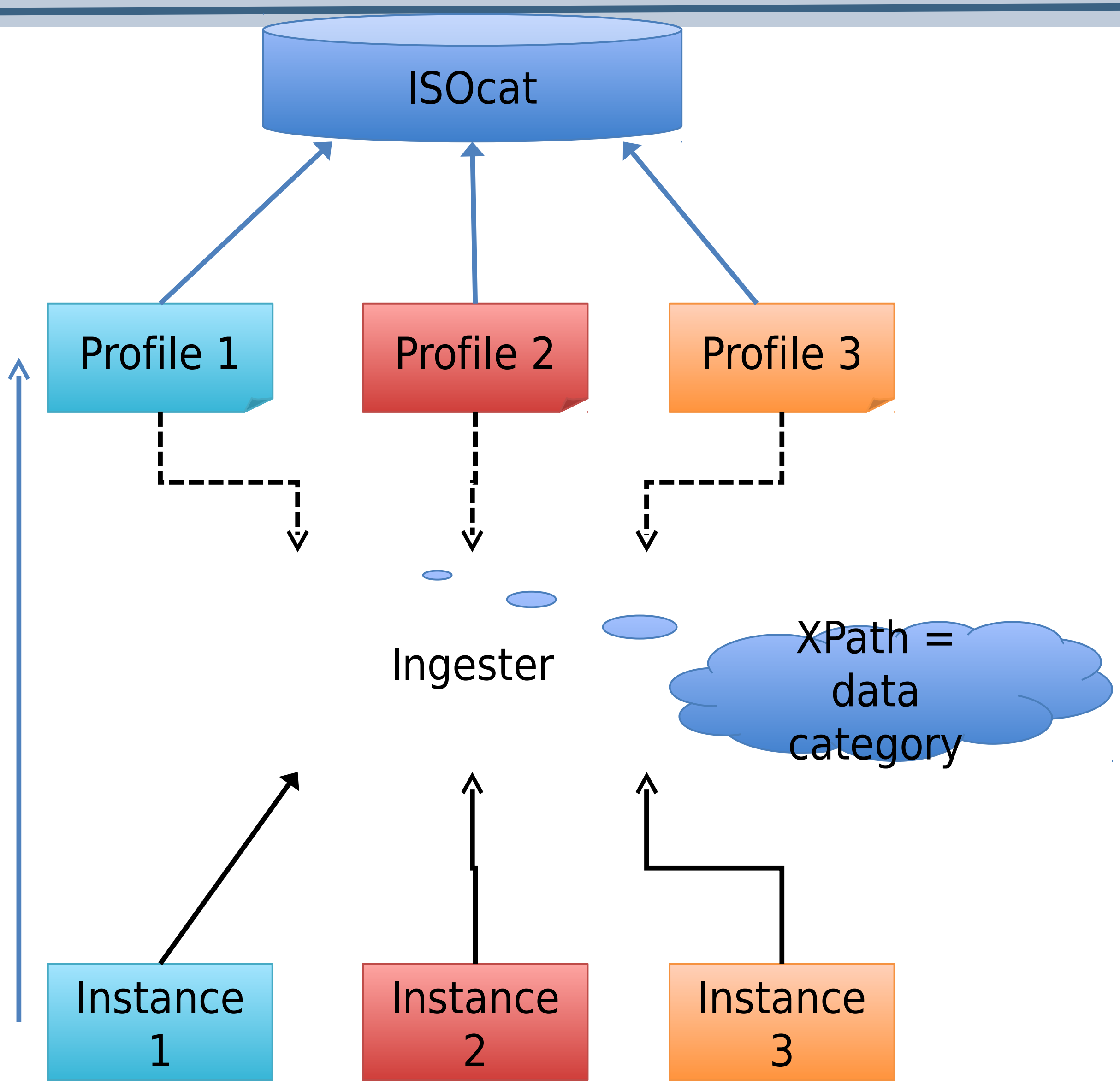
Specimen

Habitat

definitions

XSD files

CMDI files



ISOcat.org

Component registry

VLO

Metadata Repository

Still to come…

- A faceted browser is as good as its data, so curation steps are needed
- Add more CMDI metadata
- Add some more facets e.g.: year
- Human-readable hdl links
- Descriptions in the record listing
- Interface improvements
- Links to language information