CLARIN-D

The Virtual Language Observatory

Alexander König

CMDI workshop, Utrecht

2014-03-04

Overview

- VLO?
- What is behind it? – Relation to CMDI?
- How do I get my data in there?
- Demo + exercises

Context sketch

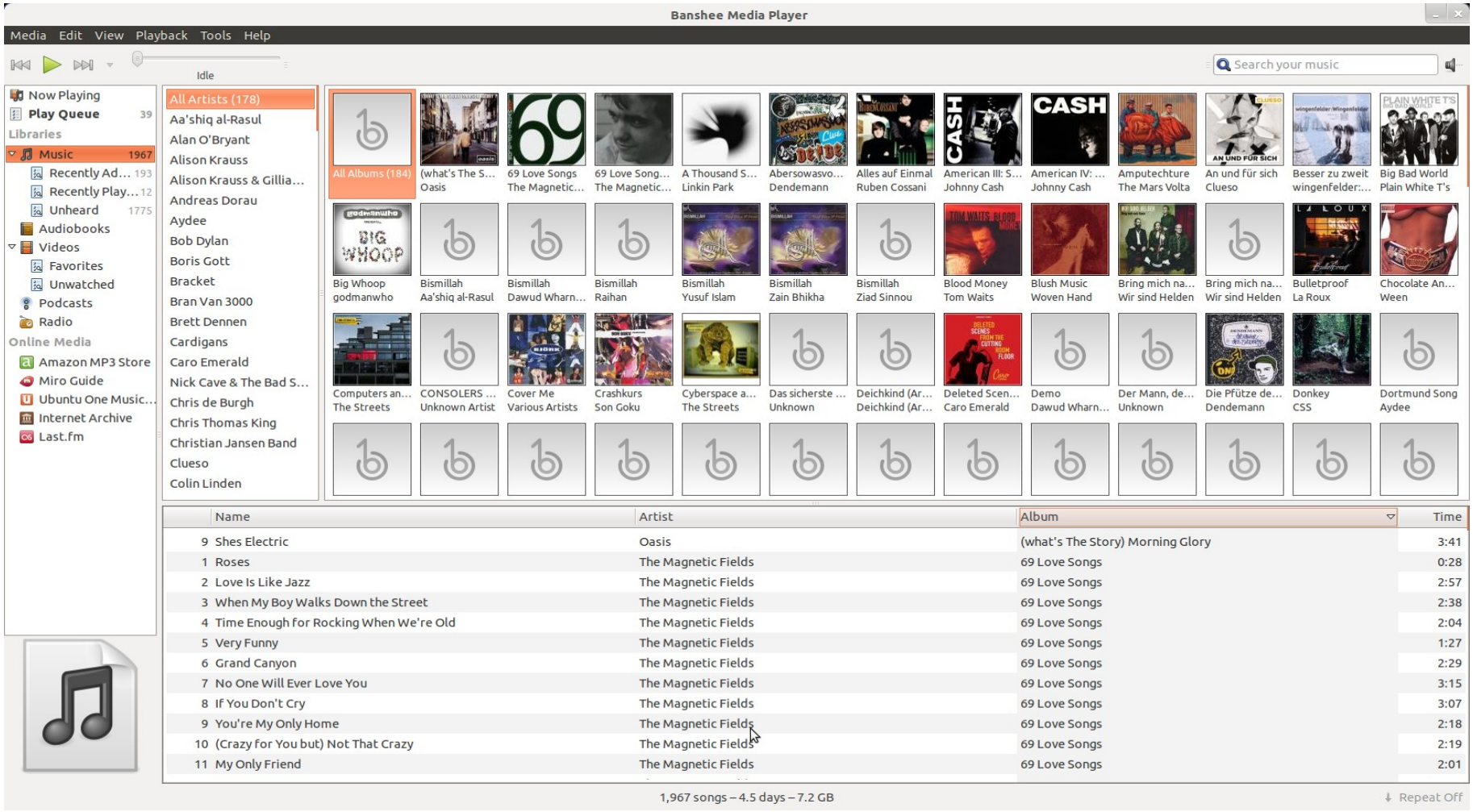
- Lots of resources somewhere out there:
  - Data collections
    - Corpora
    - Lexica
    - Grammars
    - Multimedia recordings
  - Software
  - Web applications / services
  - Old-school linguistic resources:
    - Books
    - Articles
    - CD-ROMs
- It's like a jungle, sometimes ...

Metadata to the rescue

- Metadata is everywhere

- Helping us to find things
  - A specific photo in our collection of thousands and thousands
  - The book we're looking for on Amazon
  - All Polish hip hop songs in our music collection

Metadata for music

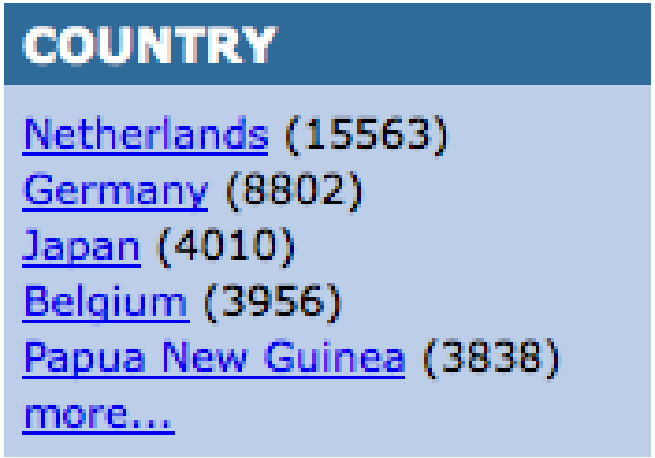


VLO: the idea

- Researcher: “where do I start?”
- Provide a single entry point giving access to all information
- Because of the large amount of data:
  - Drill-down paradigm (decrease search space gradually)
  - Multiple ways of exploring:
    - Full-text search
    - Facet browsing
  - Unified interface, links to the original context
  - http://catalog.clarin.eu/vlo

Facets?

- A **simple** way to narrow down the search space, step by step
- Values offered are **dynamic**: they change with every previous selection made
- Purpose: **quickly navigating** through a huge amount of metadata



Facets?

- Purpose: **quickly navigating** through a huge amount of resources
- Gives an idea of what is “out there”
- Useful too for **metadata curation**
- **Not** the tool to answer research questions!

VLO Faceted Browser (1)

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

VLO Faceted Browser (2)

CLARIN-D

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)

- Metadata analyzed is CMDI format
- Metadata sources
  - CMDI files harvested from CLARIN centres
  - CMDI'fied OLAC records (from CLARIN centres and others)
  - CMDI'fied LRT inventory records
- You can get to resources directly from search results

Exercises (1)

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

- **Find some resources in the catalogue:**
  - Corpus Gysseling
  - Telephone conversation recordings in Nepal

Exercises (2)

- **Find some resources in the DoBeS archive of endangered languages which are:**
  - (Spoken) discourse with at least two consultants in Asia
  - Or (spoken) discourse with at least two consultants in a Face to Face conversation

Limits

- Inherent limit: Simple search
  - no OR combinations possible
  - no sophisticated search operations

Recent Additions

- Links to language information: WALS, Wikipedia, Ethnologue, LinguistList … and the VLO
- Descriptions in the record listing
- National Project facet
- Feedback link

Still to come…

- A faceted browser is as good as its data, so further curation steps are needed
- more CMDI metadata
- some more facets e.g.: year
- Interface improvements

Questions?

- … ask them now
- … or send a mail to vlw@clarin.eu

- More information:
  - www.clarin.eu/vlo
  - www.clarin.eu/cmdi