CLARIN for Linguists Portal & Searching for Resources

Jan Odijk

LOT Summerschool

Nijmegen, 2014-06-23

Overview

- CLARIN Portal
- Find data and tools

Overview

- **CLARIN Portal**
- Find data and tools

CLARIN Portal

- Finding data and tools via one portal?
  - EU Portal: http://www.clarin.eu/guest-portal
  - NL Portal under development: sneak preview
- In the meantime, for NL:
  - Use http://www.clarin.nl/node/404
  - for an overview of CLARIN-NL
    - Data
    - Applications
    - Services
    - Links to the CLARIN Europe data and services

CLARIN Portal

- EU Portal: http://www.clarin.eu/guest-portal
  - Link to Virtual Language Observatory
  - Deposit Your Resources
  - Featured Resource
  - Showcases
  - Searching through Multiple Corpora
  - Blog
  - General information on CLARIN

CLARIN-NL Portal

- CLARIN-NL Portal sneak preview
- Intro page
- [Centres]A summary of Dutch CLARIN centres
- [Find Data] A selection of tools to find relevant resources using metadata and content
- [Services] : An inventory of tools and services with faceted search
- [CLARIN recipes] to do the most often used actions.
- [Colleagues & Advice] for asking colleagues for advice

Overview

- CLARIN Portal
- **Find data and tools**

CLARIN Infrastructure

'Can find all data and Tools'

- All data are distributed over various CLARIN centres
  - How can you find them?
- Web pages and URLs change, disappear over time
  - How do we know a link is there still tomorrow?
- Searching via Google will not work
  - Too many irrelevant search results

CLARIN Infrastructure

'Can find all data and Tools'

- CLARIN approach:
  - For all data there must be **metadata**
    - Data that describe data (resource descriptions)
  - Metadata must all be in the same format
    - Component-based Metadata Infrastructure (**CMDI**) format
  - All data and metadata must be accessible via Persistent Identifiers (**PID**)
  - Metadata must be made available according to a specific protocol (**OAI-PMH**) by each centre
  - Some centres **harvest** the metadata of all centres and store them centrally
  - And make available browser and search engines
- More on this on Friday

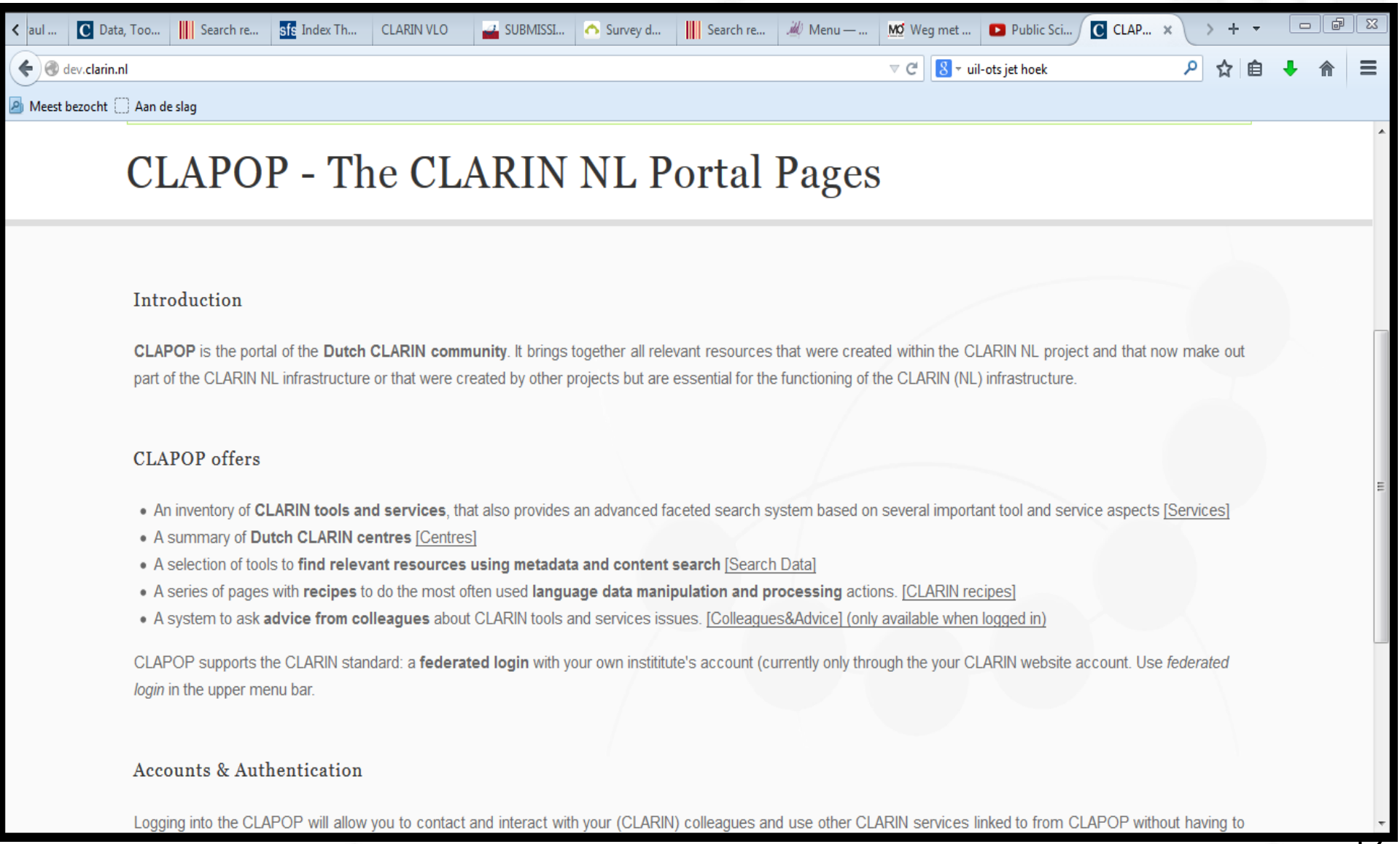
CLARIN Infrastructure

'Can find all data and Tools'

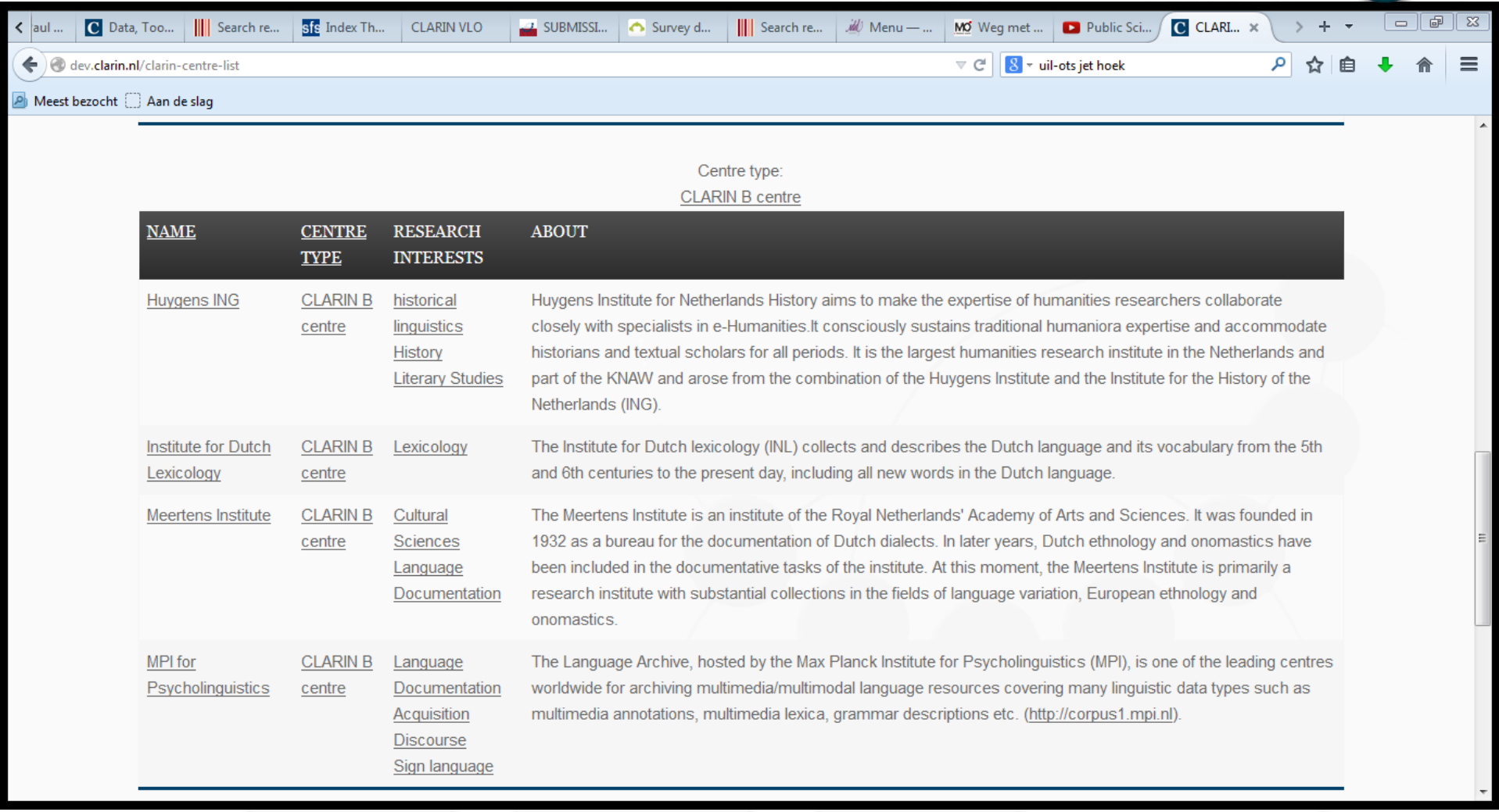
- Virtual Language Observatory
  - Faceted browsing and geographical navigation
  - CLARIN-PP
  - Demo
- Meertens CLARIN Metadata Search
  - CLARIN-NL
  - Main Features

Thanks for your Attention

CLARIN-NL Portal



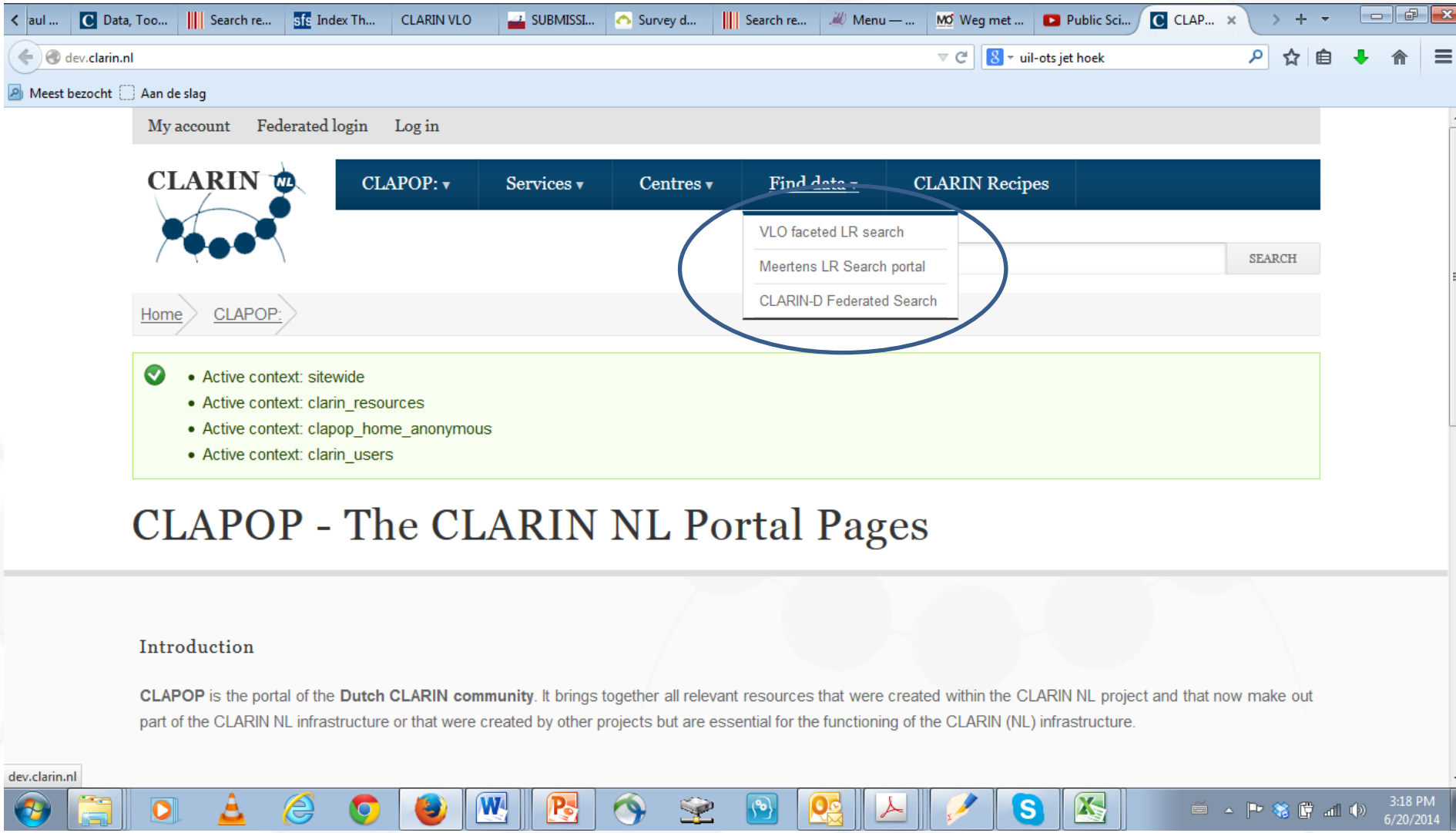
CLARIN-NL Portal



Centre type:
CLARIN B centre

NAME	CENTRE TYPE	RESEARCH INTERESTS	ABOUT
Huygens ING	CLARIN B centre	historical linguistics History Literary Studies	Huygens Institute for Netherlands History aims to make the expertise of humanities researchers collaborate closely with specialists in e-Humanities.It consciously sustains traditional humaniora expertise and accommodate historians and textual scholars for all periods. It is the largest humanities research institute in the Netherlands and part of the KNAW and arose from the combination of the Huygens Institute and the Institute for the History of the Netherlands (ING).
Institute for Dutch Lexicology	CLARIN B centre	Lexicology	The Institute for Dutch lexicology (INL) collects and describes the Dutch language and its vocabulary from the 5th and 6th centuries to the present day, including all new words in the Dutch language.
Meertens Institute	CLARIN B centre	Cultural Sciences Language Documentation	The Meertens Institute is an institute of the Royal Netherlands' Academy of Arts and Sciences. It was founded in 1932 as a bureau for the documentation of Dutch dialects. In later years, Dutch ethnology and onomastics have been included in the documentative tasks of the institute. At this moment, the Meertens Institute is primarily a research institute with substantial collections in the fields of language variation, European ethnology and onomastics.
MPI for Psycholinguistics	CLARIN B centre	Language Documentation Acquisition Discourse Sign language	The Language Archive, hosted by the Max Planck Institute for Psycholinguistics (MPI), is one of the leading centres worldwide for archiving multimedia/multimodal language resources covering many linguistic data types such as multimedia annotations, multimedia lexica, grammar descriptions etc. (http://corpus1.mpi.nl).

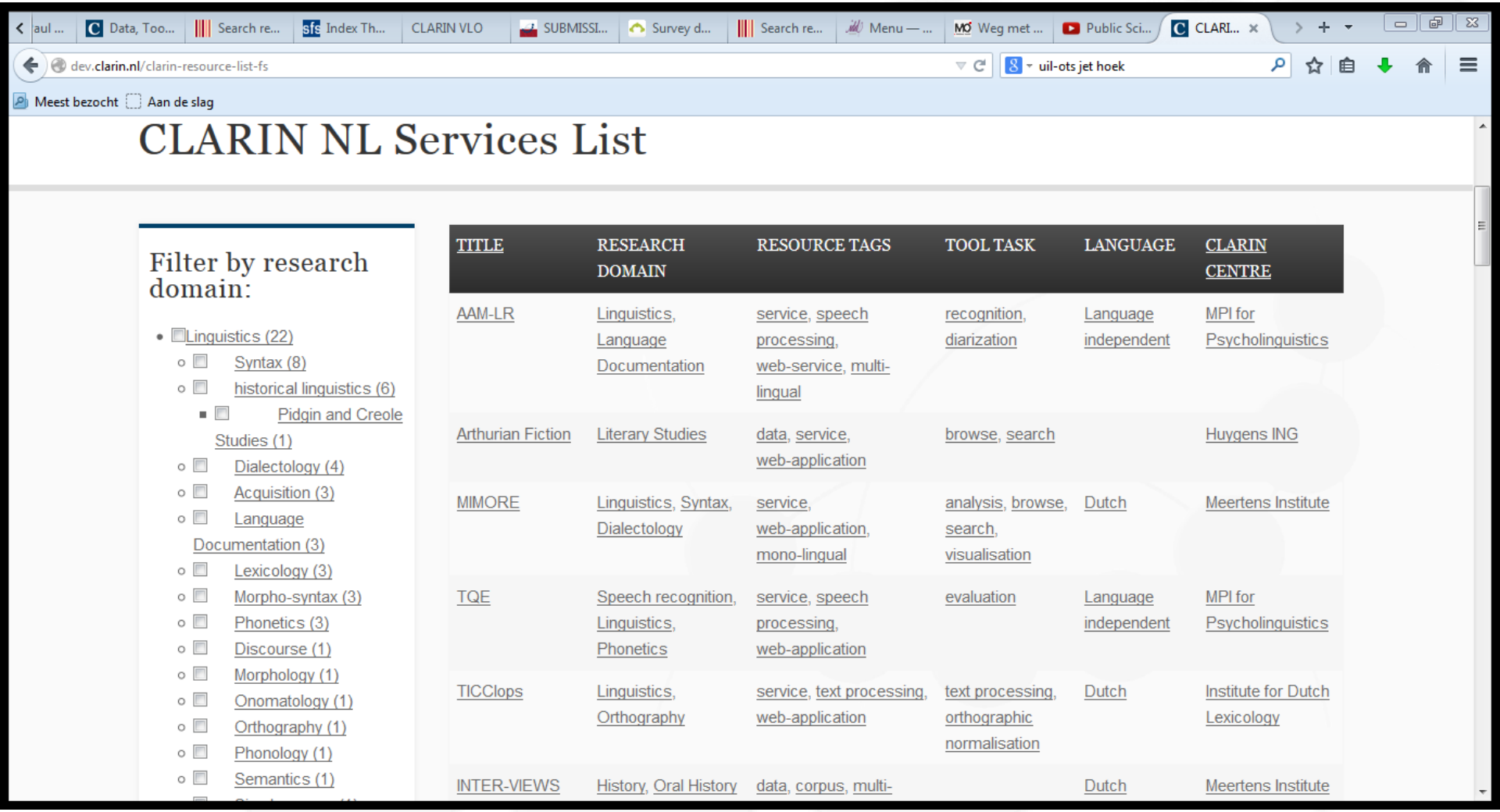
CLARIN-NL Portal



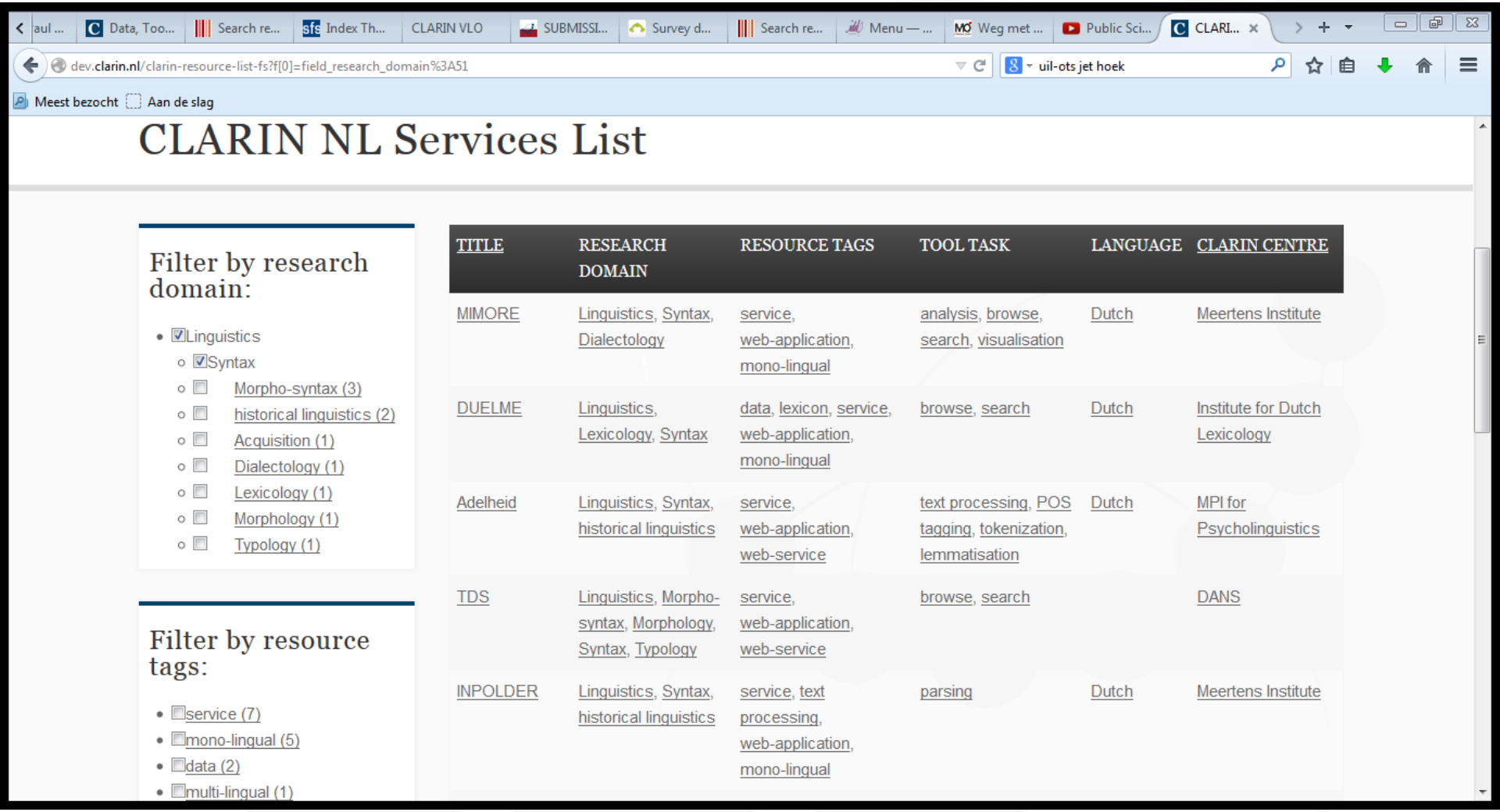
CLARIN-NL Portal

- Services List with faceted browsing
- Selection by
  - Research domain (e.g. linguistics)
  - Resource tags (e.g. data, service)
  - Tool Task (e.g. parsing, search)
  - Language (e.g. Dutch)
  - CLARIN Centre (e.g. Meertens Institute)
  - Country (e.g. Belgium)

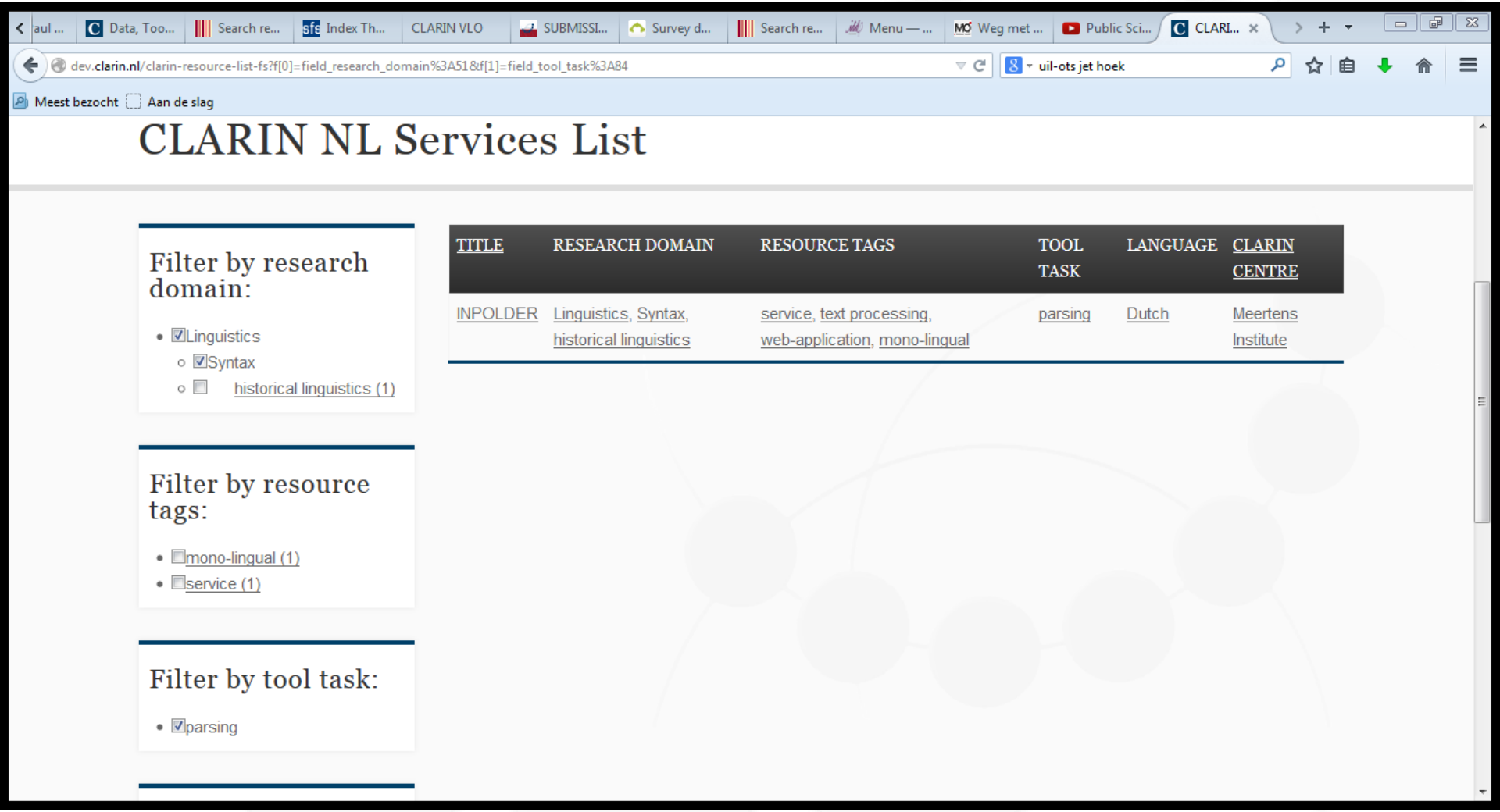
CLARIN-NL Portal



CLARIN-NL Portal



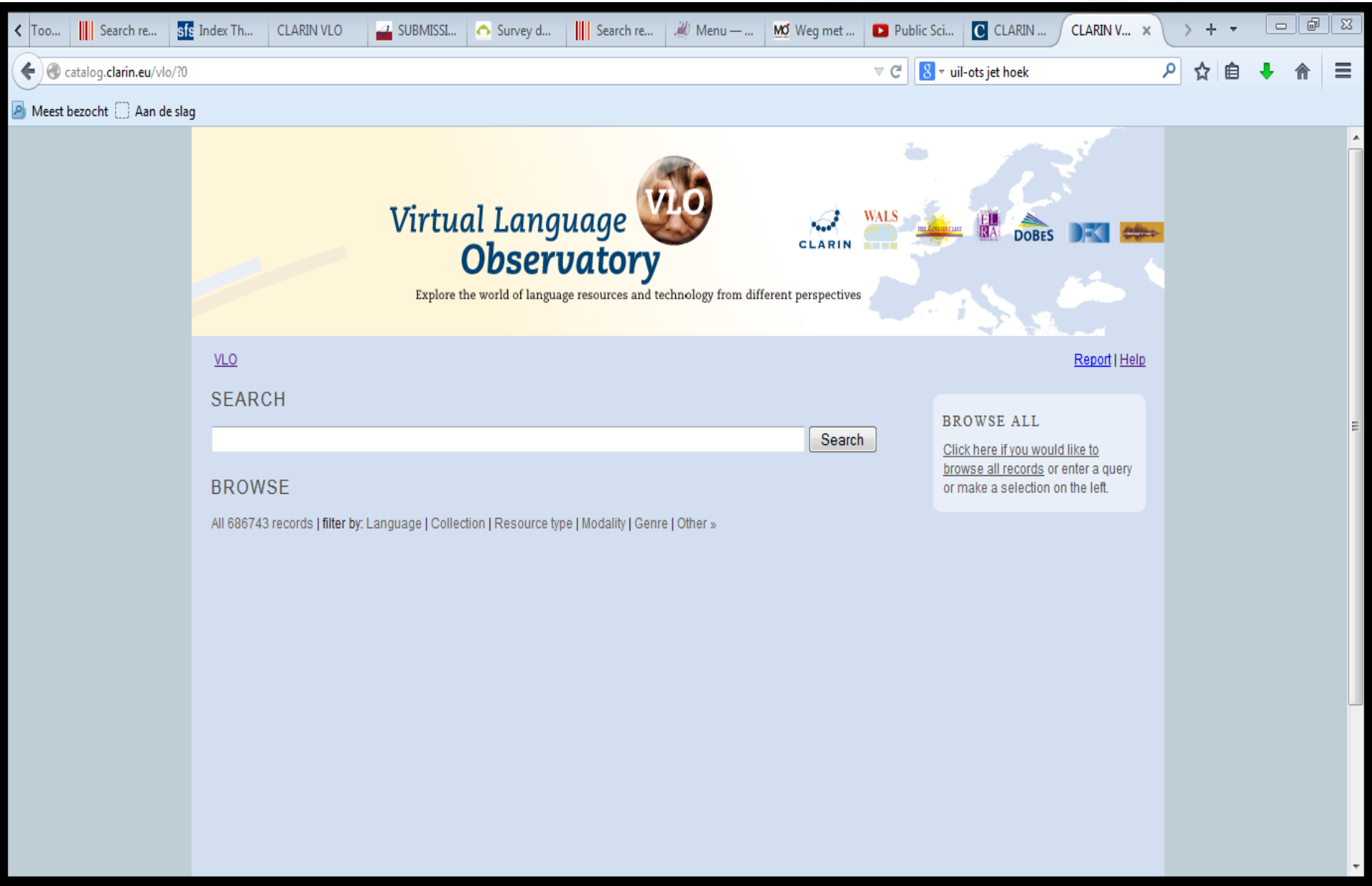
CLARIN-NL Portal



CLARIN-NL Portal

- Portal Return Page

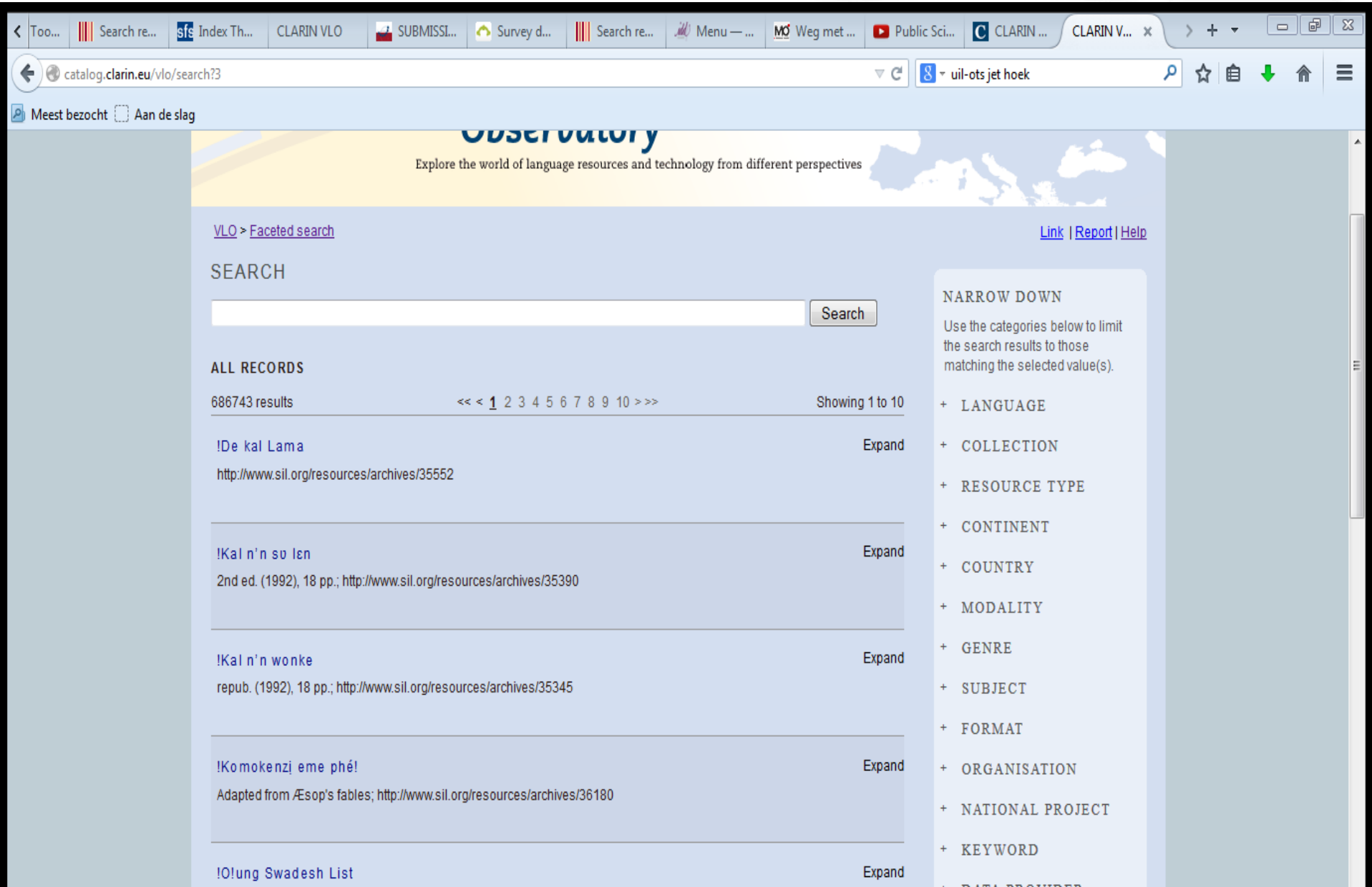
VLO


Too...
Search re...
Index Th...
CLARIN VLO
SUBMISSI...
Survey d...
Search re...
Menu — ...
Weg met ...
Public Sci...
CLARIN ...
CLARIN V...
catalog.clarin.eu/vlo/?0
uil-ots jet hoek
Meest bezocht
Aan de slag
Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DoBeS
VLO
Report | Help
SEARCH
Search
BROWSE
All 686743 records | filter by: Language | Collection | Resource type | Modality | Genre | Other »
BROWSE ALL
Click here if you would like to browse all records or enter a query or make a selection on the left.

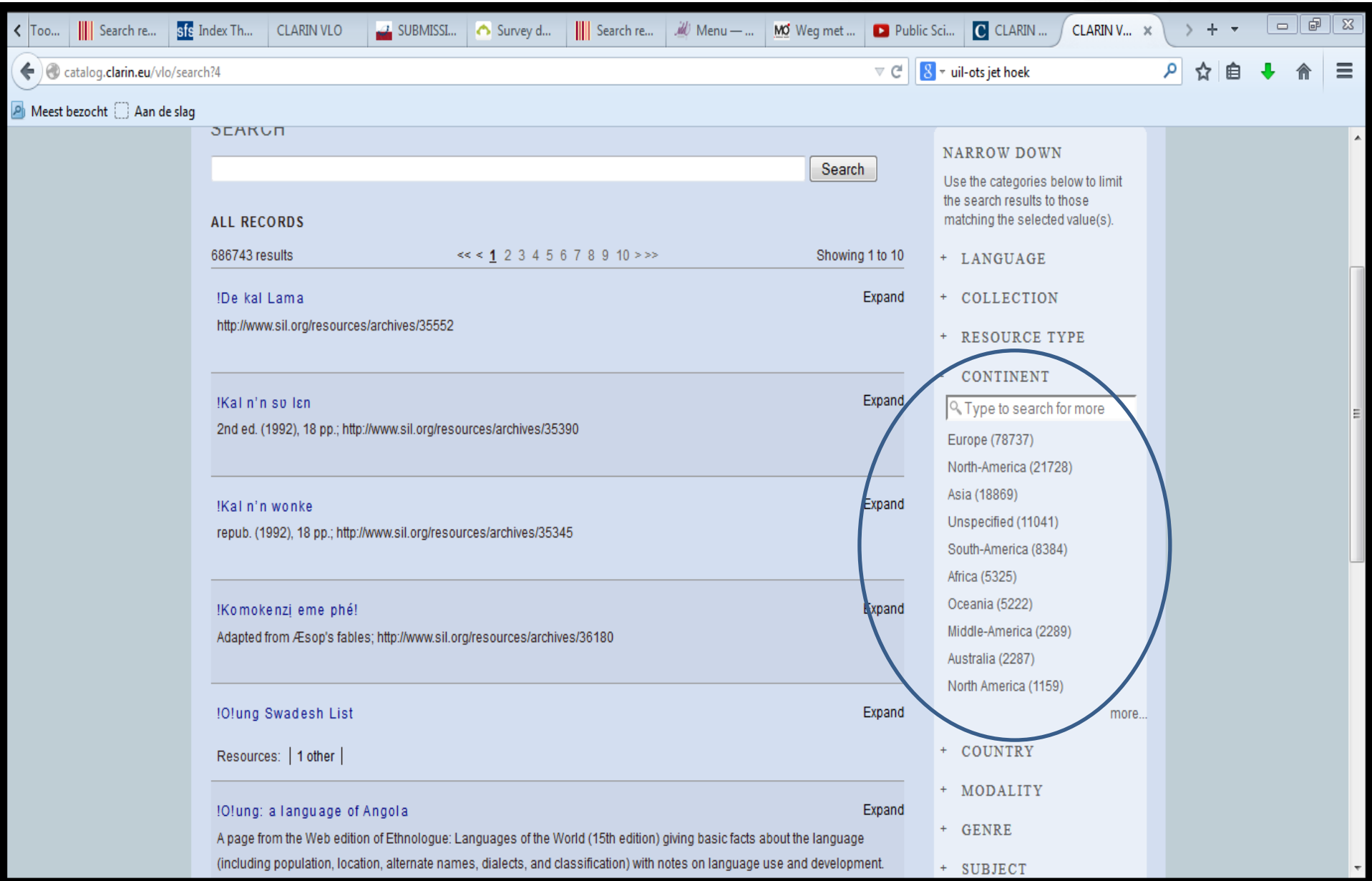
VLO

Faceted Browsing by	
Language	Subject
Collection	Format
Resource Type	Organisation
Continent	National Project
Country	Keyword
Modality	Data Provider
Genre	

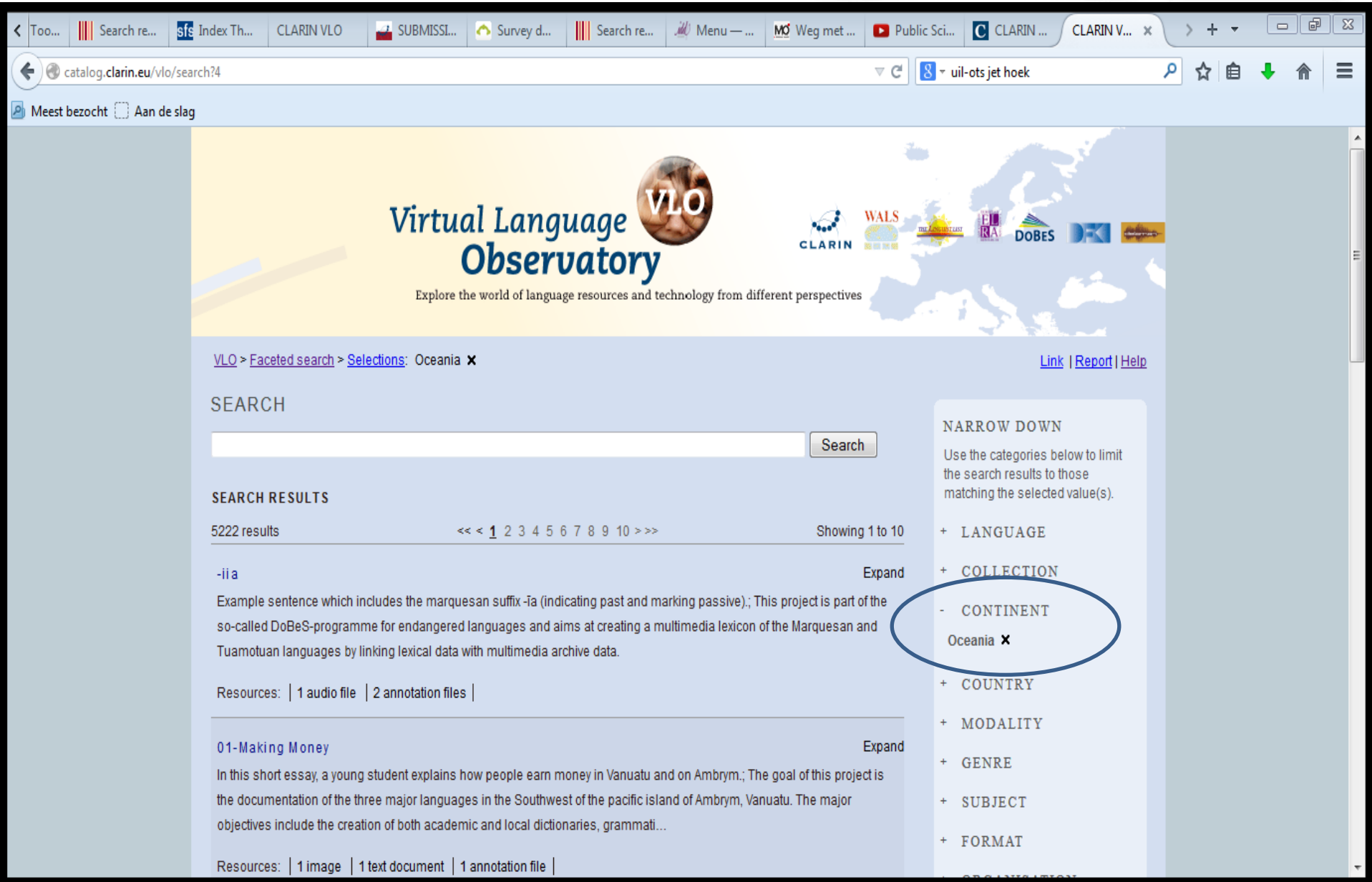
VLO


Too...
Search re...
Index Th...
CLARIN VLO
SUBMISSI...
Survey d...
Search re...
Menu — ...
Weg met ...
Public Sci...
CLARIN ...
CLARIN V...
catalog.clarin.eu/vlo/search?3
uil-ots jet hoek
Meest bezocht
Aan de slag
Explore the world of language resources and technology from different perspectives
VLO > Faceted search
Link | Report | Help
SEARCH
Search
ALL RECORDS
686743 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
!De kal Lama
Expand
http://www.sil.org/resources/archives/35552
!Kal n'n su lɛn
Expand
2nd ed. (1992), 18 pp.; http://www.sil.org/resources/archives/35390
!Kal n'n wonke
Expand
repub. (1992), 18 pp.; http://www.sil.org/resources/archives/35345
!Komokenzi eme phé!
Expand
Adapted from Æsop's fables; http://www.sil.org/resources/archives/36180
!O!ung Swadesh List
Expand
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
+ LANGUAGE
+ COLLECTION
+ RESOURCE TYPE
+ CONTINENT
+ COUNTRY
+ MODALITY
+ GENRE
+ SUBJECT
+ FORMAT
+ ORGANISATION
+ NATIONAL PROJECT
+ KEYWORD

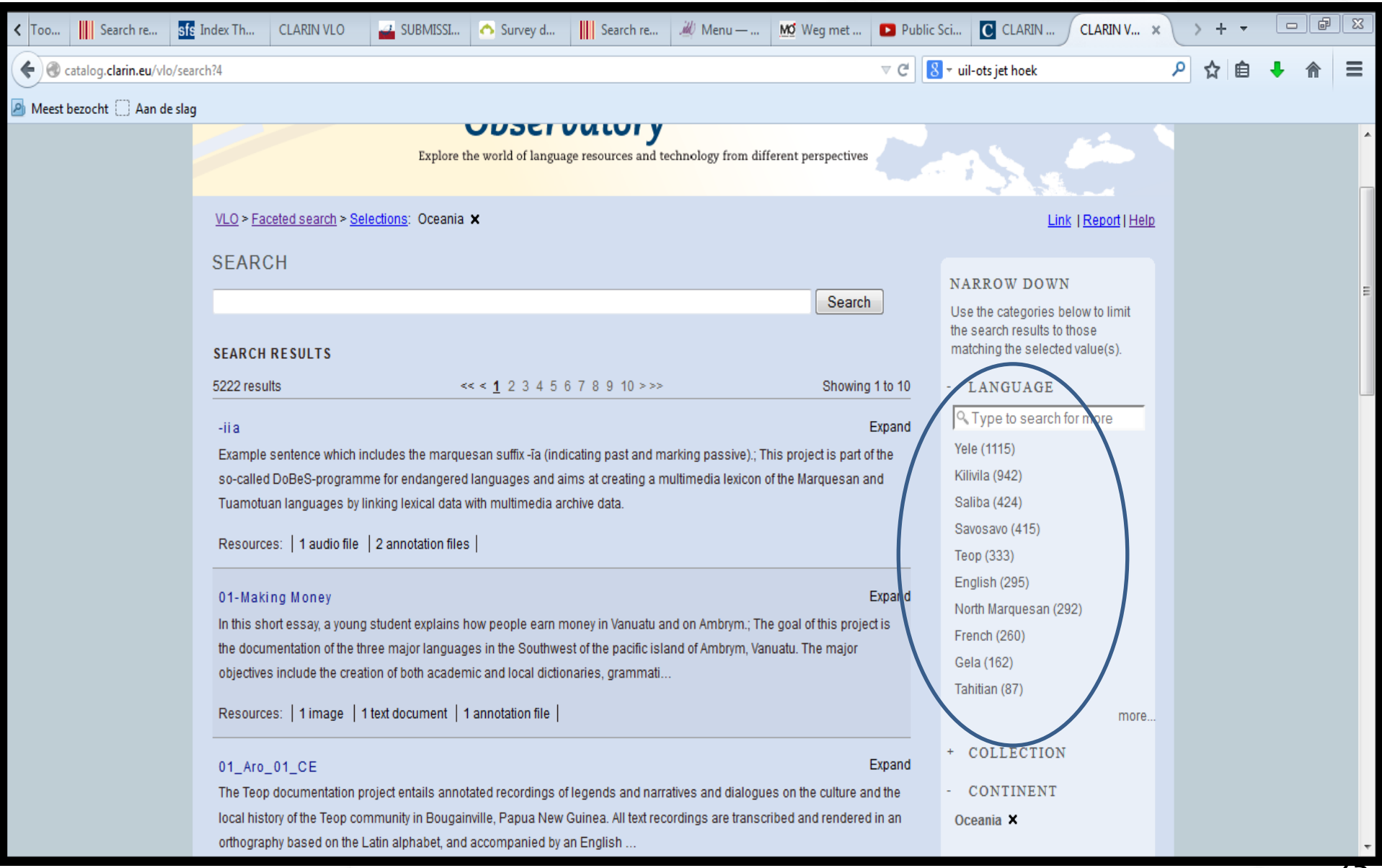
VLO


catalog.clarin.eu/vlo/search?4
uil-ots jet hoek
Meest bezocht
Aan de slag
SEARCH
Search
ALL RECORDS
686743 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
!De kal Lama
Expand
http://www.sil.org/resources/archives/35552
!Kal n'n sʊ lɛn
Expand
2nd ed. (1992), 18 pp.; http://www.sil.org/resources/archives/35390
!Kal n'n wonke
Expand
repub. (1992), 18 pp.; http://www.sil.org/resources/archives/35345
!Komokenzi eme phé!
Expand
Adapted from Æsop's fables; http://www.sil.org/resources/archives/36180
!O!ung Swadesh List
Expand
Resources: | 1 other |
!O!ung: a language of Angola
Expand
A page from the Web edition of Ethnologue: Languages of the World (15th edition) giving basic facts about the language (including population, location, alternate names, dialects, and classification) with notes on language use and development.
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
+ LANGUAGE
+ COLLECTION
+ RESOURCE TYPE
- CONTINENT
Type to search for more
Europe (78737)
North-America (21728)
Asia (18869)
Unspecified (11041)
South-America (8384)
Africa (5325)
Oceania (5222)
Middle-America (2289)
Australia (2287)
North America (1159)
more...
+ COUNTRY
+ MODALITY
+ GENRE
+ SUBJECT

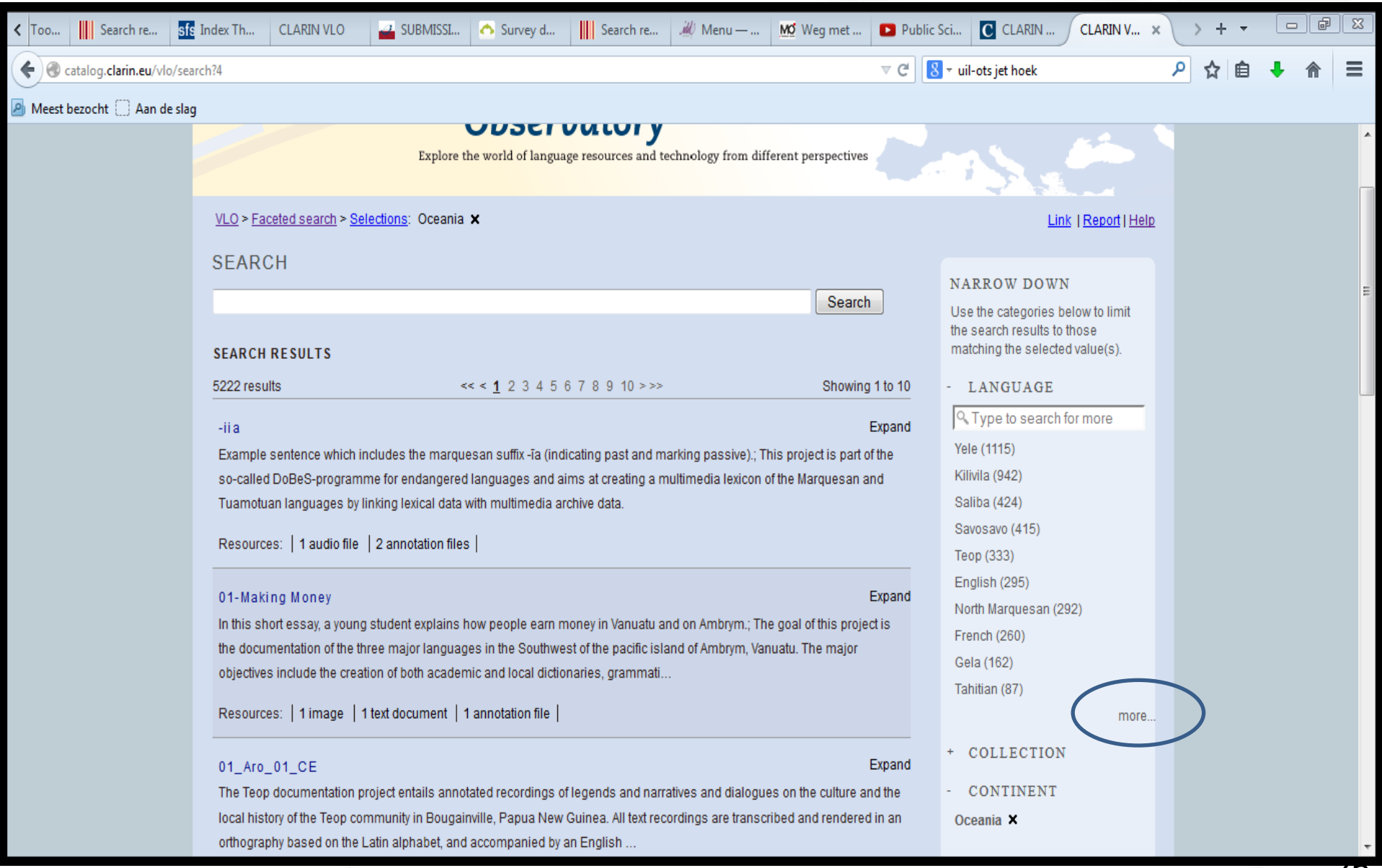
VLO


catalog.clarin.eu/vlo/search?4
uil-ots jet hoek
Meest bezocht
Aan de slag
Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DoBeS
VLO > Faceted search > Selections: Oceania ✖
Link | Report | Help
SEARCH
Search
SEARCH RESULTS
5222 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
-ii a
Expand
Example sentence which includes the marquesan suffix -īa (indicating past and marking passive).; This project is part of the so-called DoBeS-programme for endangered languages and aims at creating a multimedia lexicon of the Marquesan and Tuamotuan languages by linking lexical data with multimedia archive data.
Resources: | 1 audio file | 2 annotation files |
01-Making Money
Expand
In this short essay, a young student explains how people earn money in Vanuatu and on Ambrym.; The goal of this project is the documentation of the three major languages in the Southwest of the pacific island of Ambrym, Vanuatu. The major objectives include the creation of both academic and local dictionaries, grammati...
Resources: | 1 image | 1 text document | 1 annotation file |
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
+ LANGUAGE
+ COLLECTION
- CONTINENT
Oceania ✖
+ COUNTRY
+ MODALITY
+ GENRE
+ SUBJECT
+ FORMAT

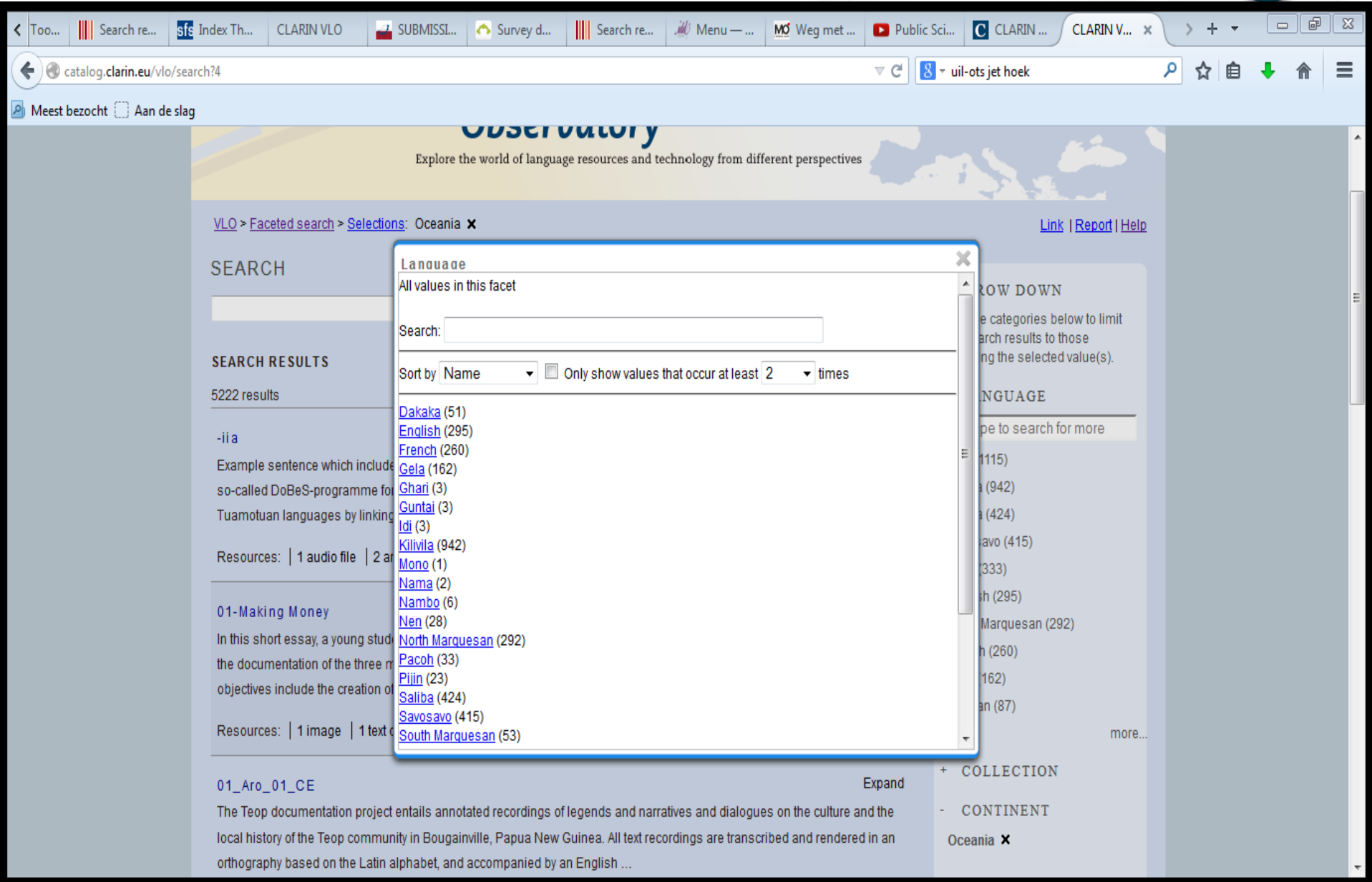
VLO


catalog.clarin.eu/vlo/search?4
uil-ots jet hoek
Meest bezocht
Aan de slag
Explore the world of language resources and technology from different perspectives
VLO > Faceted search > Selections: Oceania ✖
Link | Report | Help
SEARCH
Search
SEARCH RESULTS
5222 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
-iia
Expand
Example sentence which includes the marquesan suffix -īa (indicating past and marking passive).; This project is part of the so-called DoBeS-programme for endangered languages and aims at creating a multimedia lexicon of the Marquesan and Tuamotuan languages by linking lexical data with multimedia archive data.
Resources: | 1 audio file | 2 annotation files |
01-Making Money
Expand
In this short essay, a young student explains how people earn money in Vanuatu and on Ambrym.; The goal of this project is the documentation of the three major languages in the Southwest of the pacific island of Ambrym, Vanuatu. The major objectives include the creation of both academic and local dictionaries, grammati...
Resources: | 1 image | 1 text document | 1 annotation file |
01_Aro_01_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
- LANGUAGE
Type to search for more
Yele (1115)
Kilivila (942)
Saliba (424)
Savosavo (415)
Teop (333)
English (295)
North Marquesan (292)
French (260)
Gela (162)
Tahitian (87)
more...
+ COLLECTION
- CONTINENT
Oceania ✖

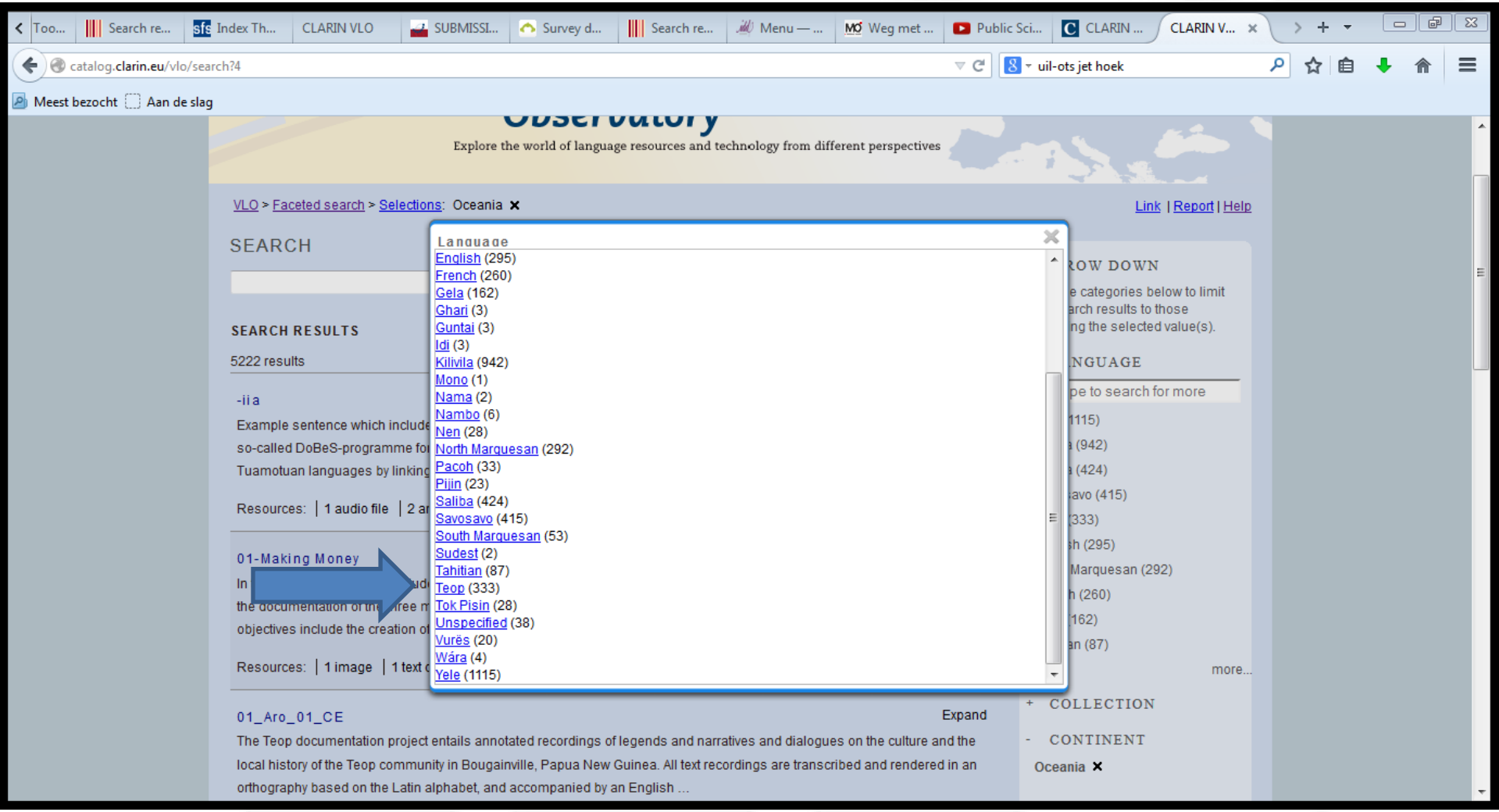
VLO


catalog.clarin.eu/vlo/search?4
uil-ots jet hoek
Meest bezocht
Aan de slag
Explore the world of language resources and technology from different perspectives
VLO > Faceted search > Selections: Oceania ✖
Link | Report | Help
SEARCH
Search
SEARCH RESULTS
5222 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
-iia
Expand
Example sentence which includes the marquesan suffix -īa (indicating past and marking passive).; This project is part of the so-called DoBeS-programme for endangered languages and aims at creating a multimedia lexicon of the Marquesan and Tuamotuan languages by linking lexical data with multimedia archive data.
Resources: | 1 audio file | 2 annotation files |
01-Making Money
Expand
In this short essay, a young student explains how people earn money in Vanuatu and on Ambrym.; The goal of this project is the documentation of the three major languages in the Southwest of the pacific island of Ambrym, Vanuatu. The major objectives include the creation of both academic and local dictionaries, grammati...
Resources: | 1 image | 1 text document | 1 annotation file |
01_Aro_01_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
- LANGUAGE
Type to search for more
Yele (1115)
Kilivila (942)
Saliba (424)
Savosavo (415)
Teop (333)
English (295)
North Marquesan (292)
French (260)
Gela (162)
Tahitian (87)
more...
+ COLLECTION
- CONTINENT
Oceania ✖

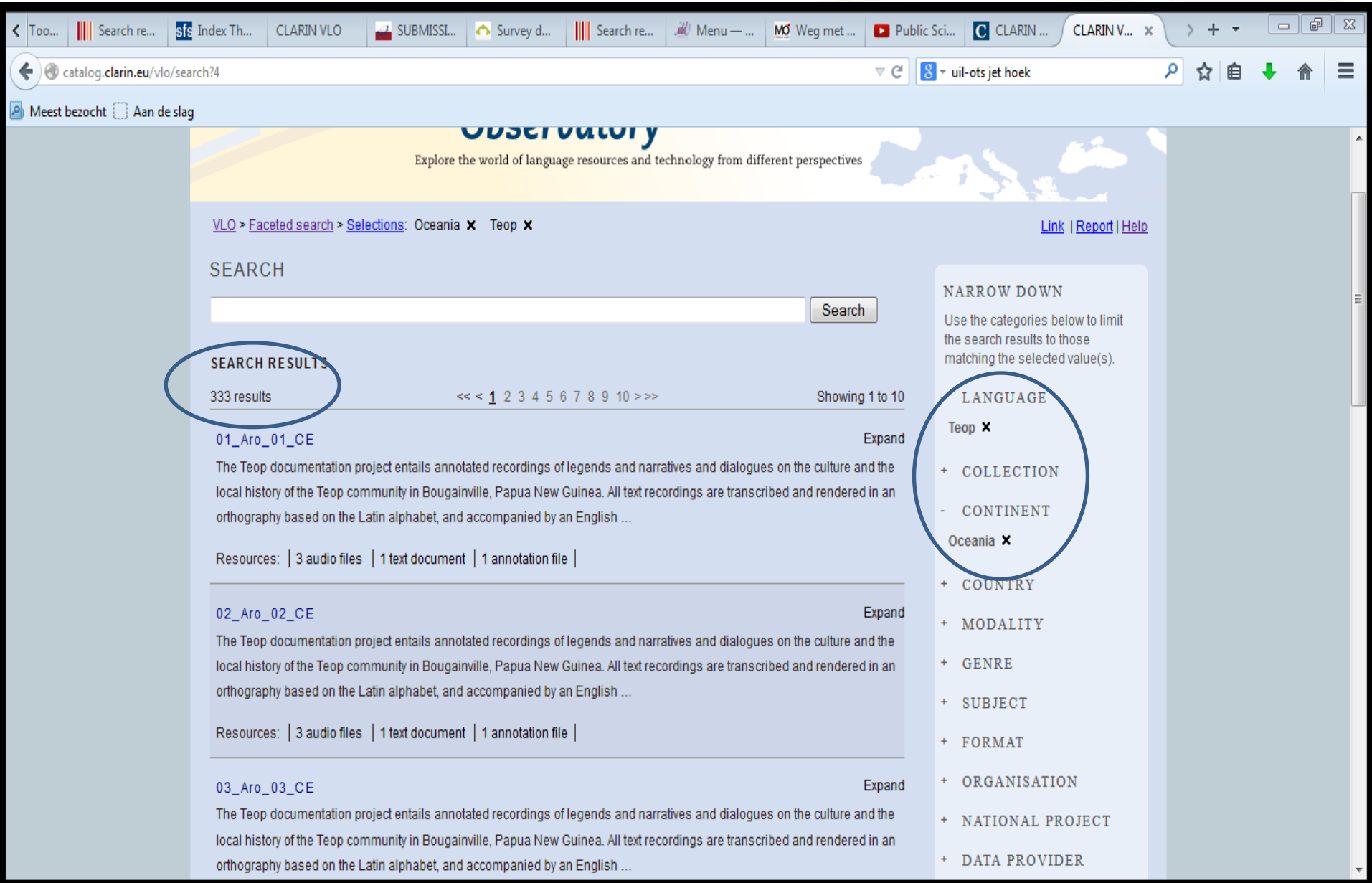
VLO


catalog.clarin.eu/vlo/search?4
uil-ots jet hoek
Meest bezocht
Aan de slag
Explore the world of language resources and technology from different perspectives
VLO > Faceted search > Selections: Oceania ✖
Link | Report | Help
SEARCH
SEARCH RESULTS
5222 results
Language
All values in this facet
Search:
Sort by Name
Only show values that occur at least 2 times
Dakaka (51)
English (295)
French (260)
Gela (162)
Ghari (3)
Guntai (3)
Idi (3)
Kilivila (942)
Mono (1)
Nama (2)
Nambo (6)
Nen (28)
North Marquesan (292)
Pacoh (33)
Pijin (23)
Saliba (424)
Savosavo (415)
South Marquesan (53)
01_Aro_01_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
+ COLLECTION
- CONTINENT
Oceania ✖
more...

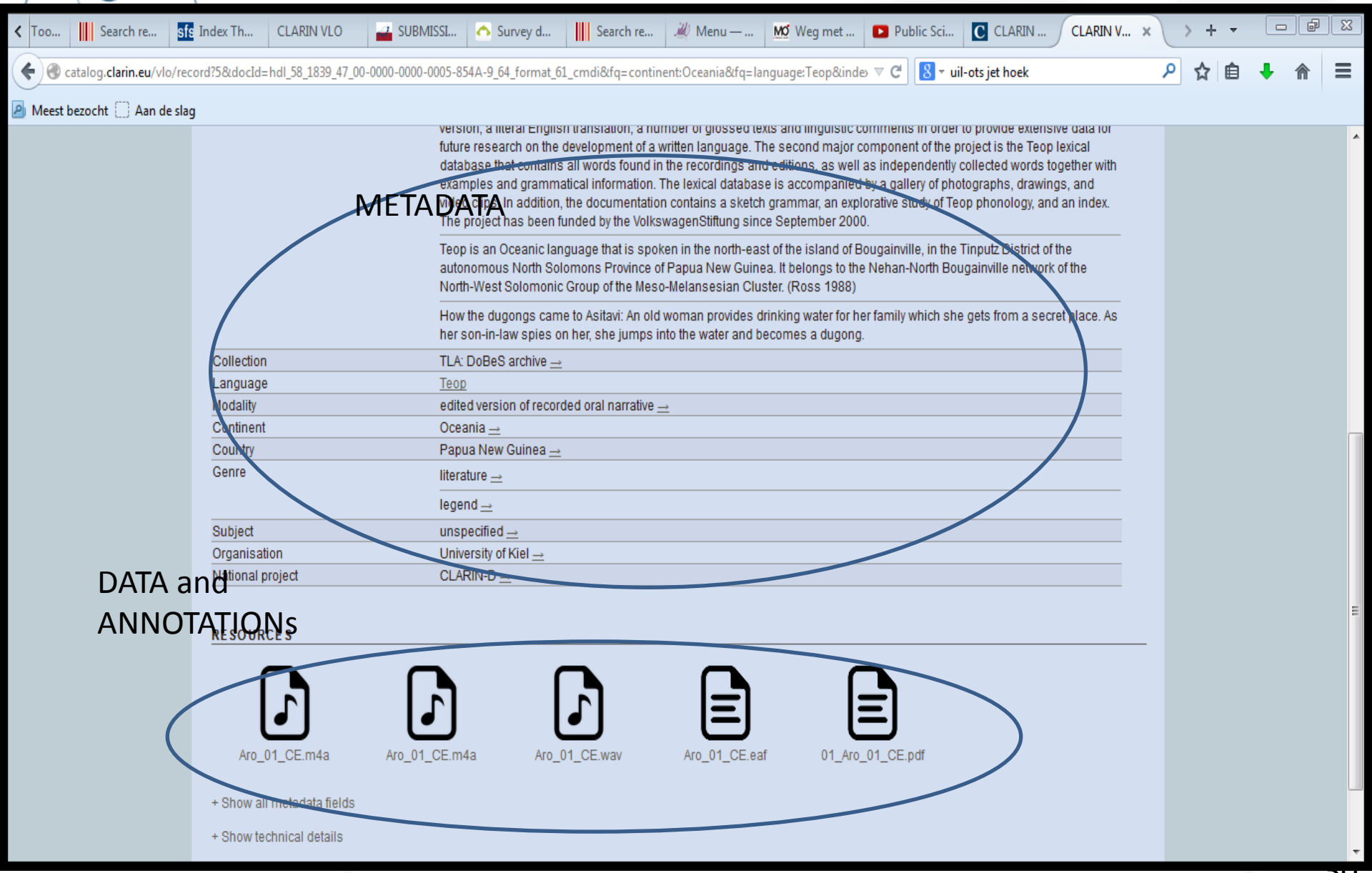
VLO


catalog.clarin.eu/vlo/search?4
Explore the world of language resources and technology from different perspectives
VLO > Faceted search > Selections: Oceania ✖
Link | Report | Help
SEARCH
SEARCH RESULTS
5222 results
Language
English (295)
French (260)
Gela (162)
Ghari (3)
Guntai (3)
Idi (3)
Kilivila (942)
Mono (1)
Nama (2)
Nambo (6)
Nen (28)
North Marquesan (292)
Pacoh (33)
Pijin (23)
Saliba (424)
Savosavo (415)
South Marquesan (53)
Sudest (2)
Tahitian (87)
Teop (333)
Tok Pisin (28)
Unspecified (38)
Vurës (20)
Wára (4)
Yele (1115)
01-Making Money
01_Aro_01_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
COLLECTION
CONTINENT
Oceania ✖

VLO


catalog.clarin.eu/vlo/search?4
Explore the world of language resources and technology from different perspectives
VLO > Faceted search > Selections: Oceania ✖ Teop ✖
Link | Report | Help
SEARCH
Search
SEARCH RESULTS
333 results
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Showing 1 to 10
01_Aro_01_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
Resources: | 3 audio files | 1 text document | 1 annotation file |
02_Aro_02_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
Resources: | 3 audio files | 1 text document | 1 annotation file |
03_Aro_03_CE
Expand
The Teop documentation project entails annotated recordings of legends and narratives and dialogues on the culture and the local history of the Teop community in Bougainville, Papua New Guinea. All text recordings are transcribed and rendered in an orthography based on the Latin alphabet, and accompanied by an English ...
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
LANGUAGE
Teop ✖
+ COLLECTION
- CONTINENT
Oceania ✖
+ COUNTRY
+ MODALITY
+ GENRE
+ SUBJECT
+ FORMAT
+ ORGANISATION
+ NATIONAL PROJECT
+ DATA PROVIDER

VLO


catalog.clarin.eu/vlo/record?5&docId=hdl_58_1839_47_00-0000-0000-0005-854A-9_64_format_61_cmdi&fq=continent:Oceania&fq=language:Teop&inde
uil-ots jet hoek
Meest bezocht
Aan de slag
METADATA
DATA and ANNOTATIONs
Collection
TLA: DoBeS archive
Language
Teop
Modality
edited version of recorded oral narrative
Continent
Oceania
Country
Papua New Guinea
Genre
literature
legend
Subject
unspecified
Organisation
University of Kiel
National project
CLARIN-D
RESOURCES
Aro_01_CE.m4a
Aro_01_CE.m4a
Aro_01_CE.wav
Aro_01_CE.eaf
01_Aro_01_CE.pdf
+ Show all metadata fields
+ Show technical details

VLO

- VLO Return Page

Meertens Metadata Search

- **Query Completion**
  - automatic query completion + hit count + tag
- Suggestions for other searches
  - You could also look for...
- Storing the search selection
- Different Overviews of the Retrieved results
  - dynamic tag cloud of the aggregated content within the metadata label

Meertens Metadata Search

- Visualisation of aggregated search features
  - Geo-referenced resources on a map
  - Editable charts for displaying the date ranges of documents
- Recommend related resources
  - link to the metadata record
  - snippet of a recommendation

Meertens Metadata Search

- Meertens Return Page

Meertens Metadata Search

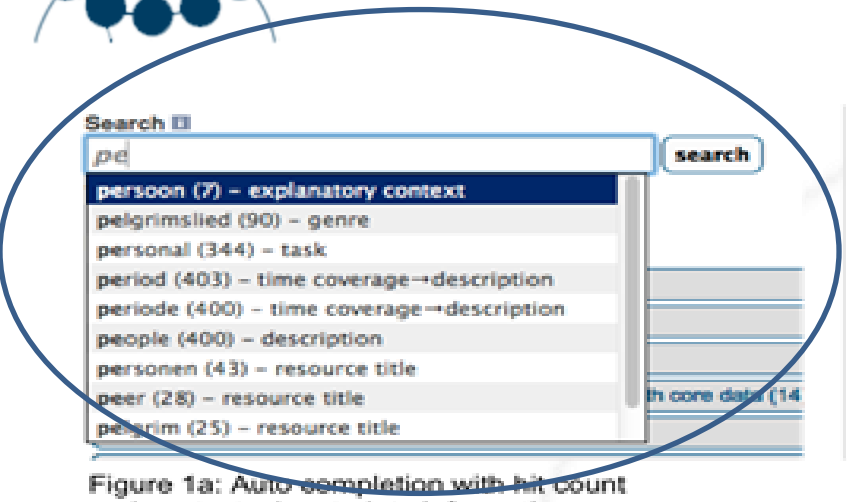


Figure 1a: Auto-completion with hit count and contextual metadata information

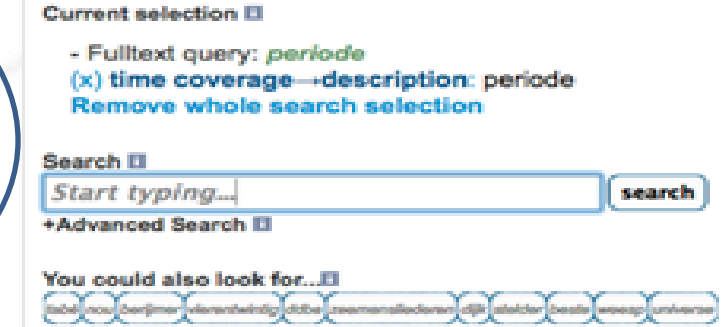


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

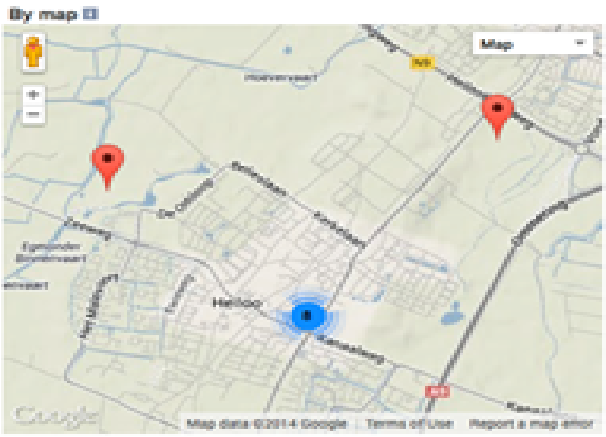


Figure 1c: Tag cloud distribution and geo referenced map distribution of search results

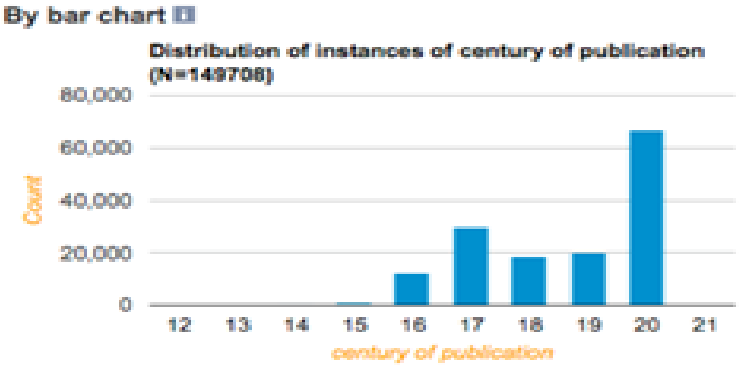


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

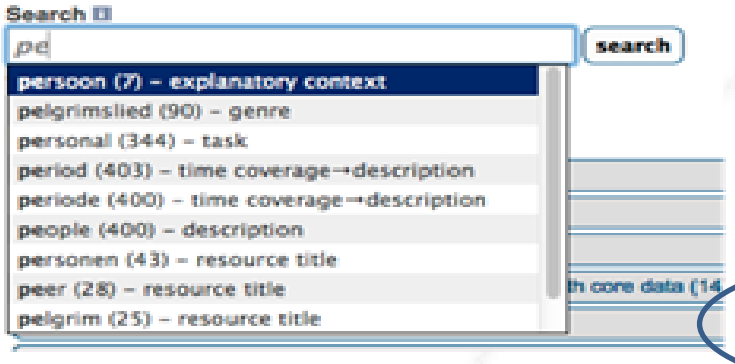


Figure 1a: Auto completion with hit count and contextual metadata information

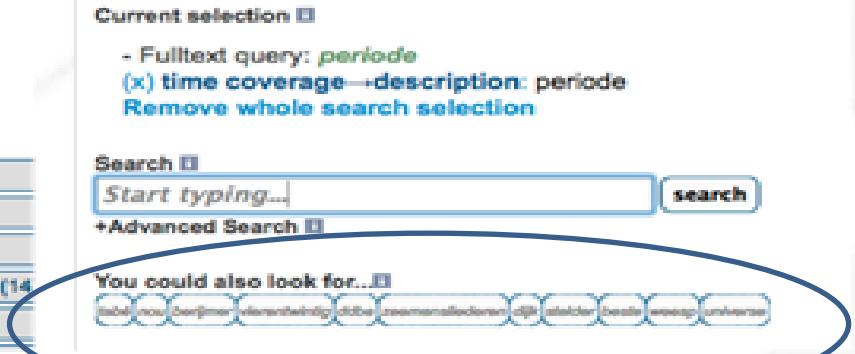


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

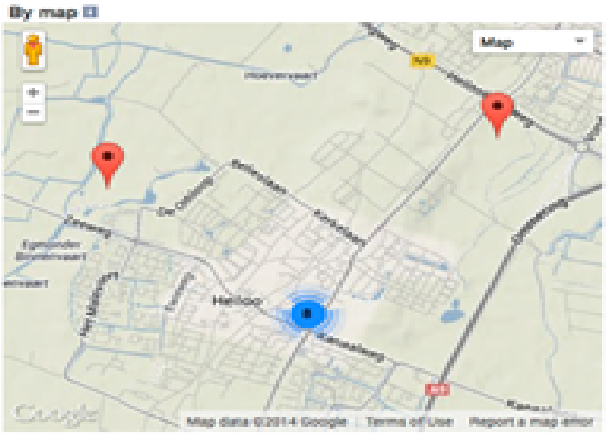


Figure 1c: Tag cloud distribution and geo referenced map distribution of search results

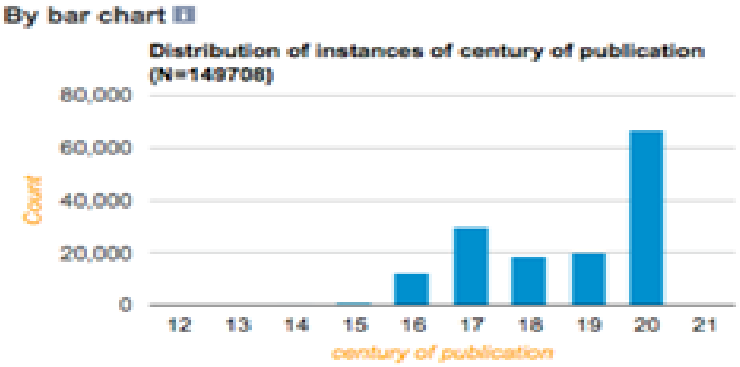


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

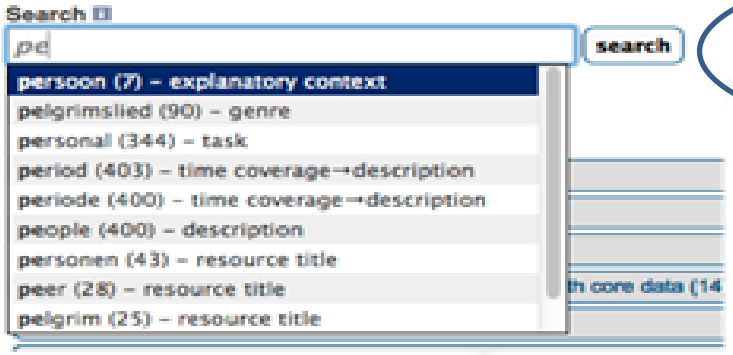


Figure 1a: Auto completion with hit count and contextual metadata information

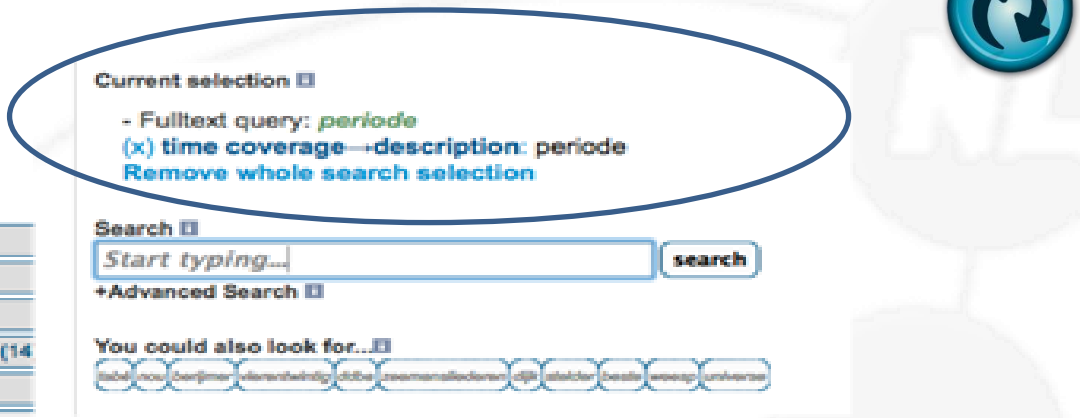


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

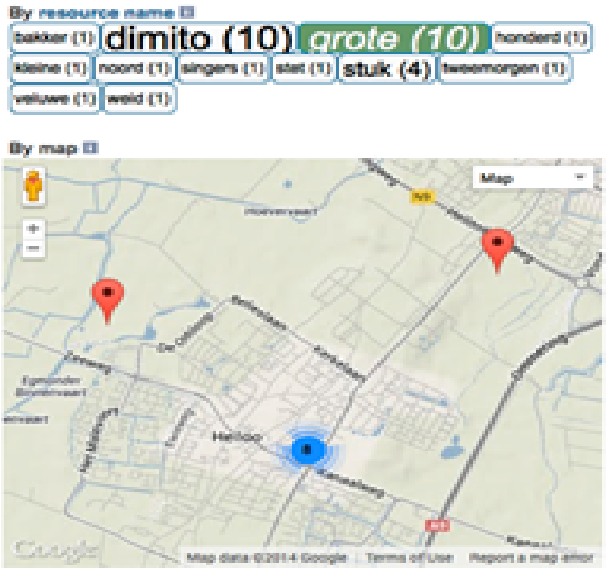


Figure 1c: Tag cloud distribution and geo referenced map distribution of search results

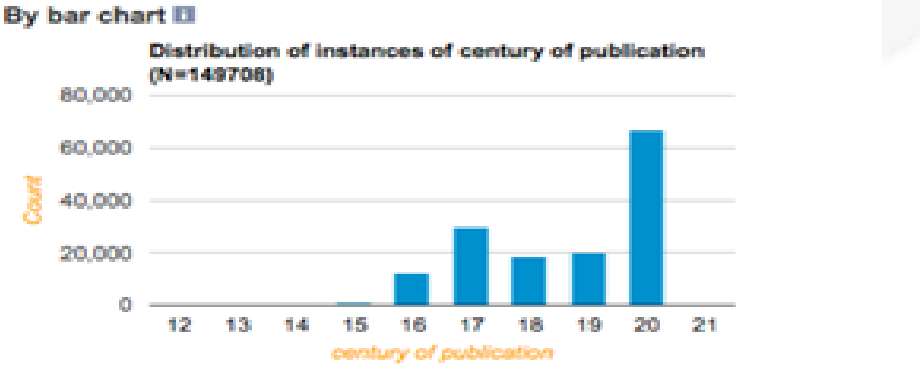


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

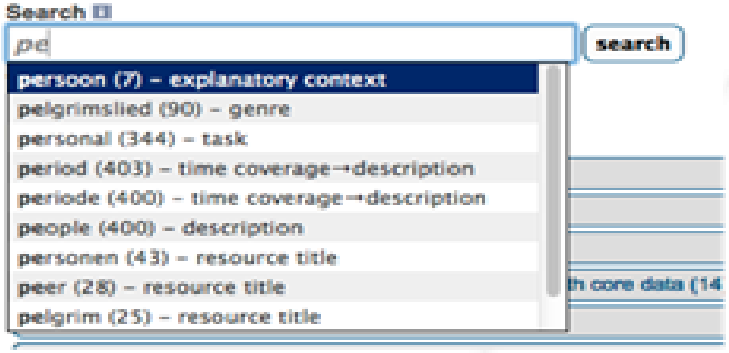


Figure 1a: Auto completion with hit count and contextual metadata information

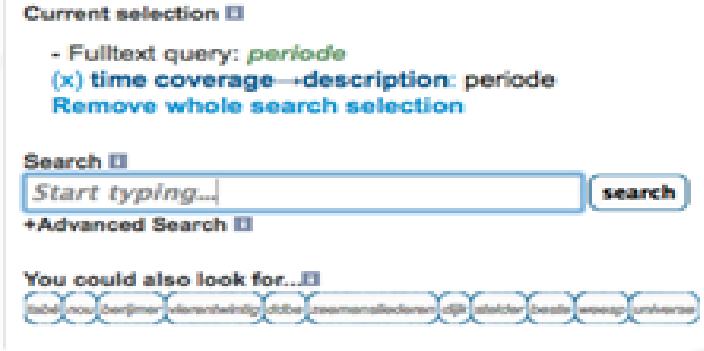


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

Figure 1c: Tag cloud distribution and geo referenced map distribution of search results

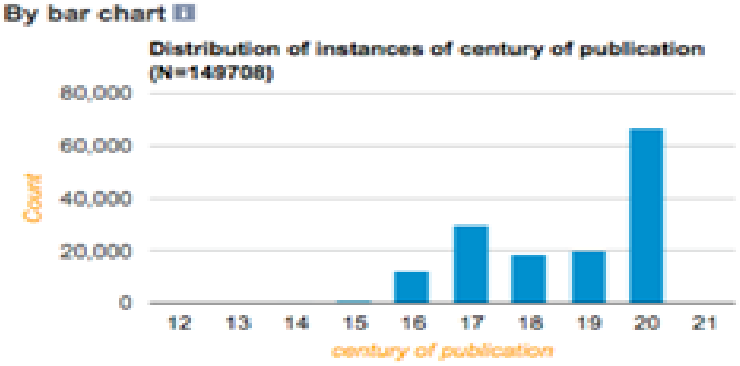


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

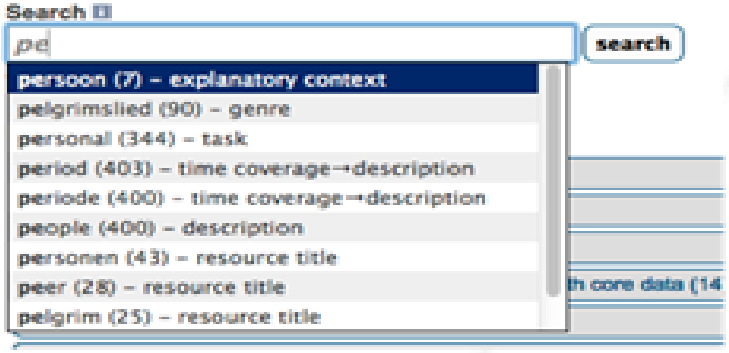


Figure 1a: Auto completion with hit count and contextual metadata information

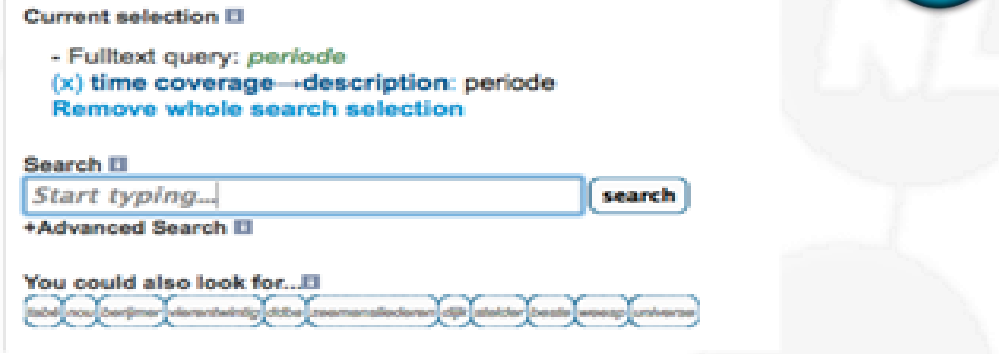


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

Figure 1c: Tag cloud distribution and geo-referenced map distribution of search results

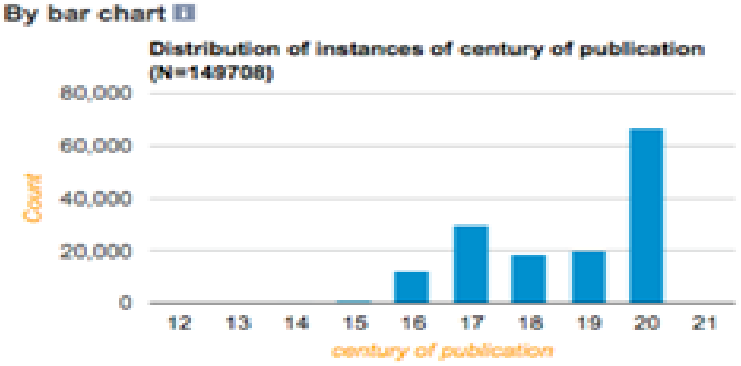


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

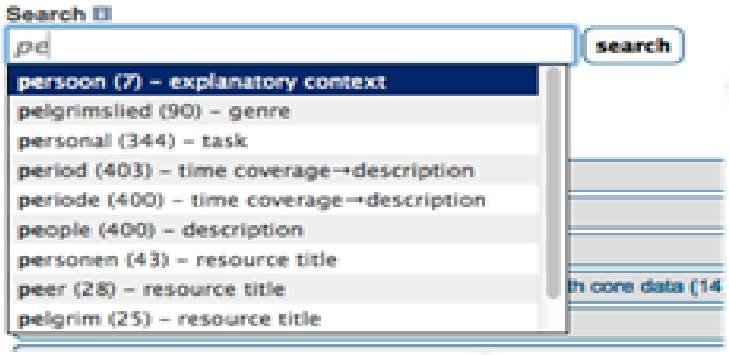


Figure 1a: Auto completion with hit count and contextual metadata information

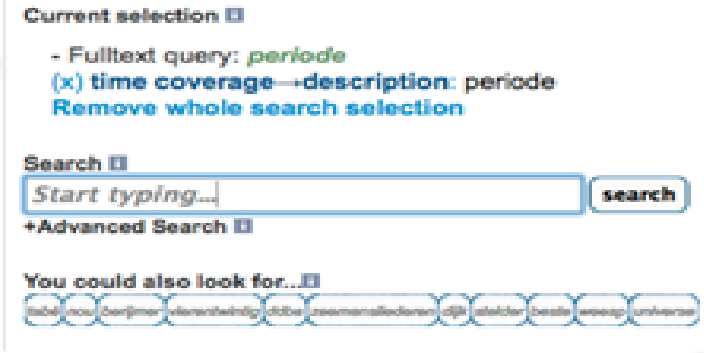


Figure 1b: Query history widget with query and metadata context information. Related terms are presented using the top TF*IDF terms.

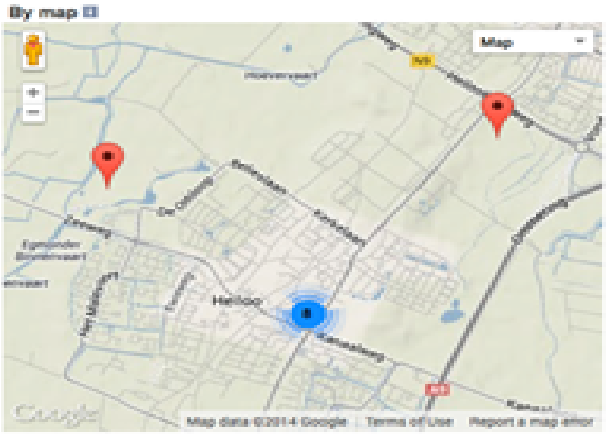


Figure 1c: Tag cloud distribution and geo referenced map distribution of search results

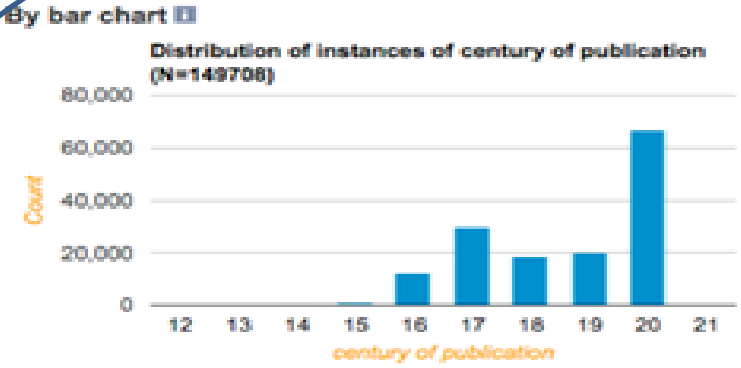


Figure 1d: Bar chart distribution for time referenced search results

Meertens Metadata Search

Figure 2a: Customized views different CMDI profiles displaying relevant profile information

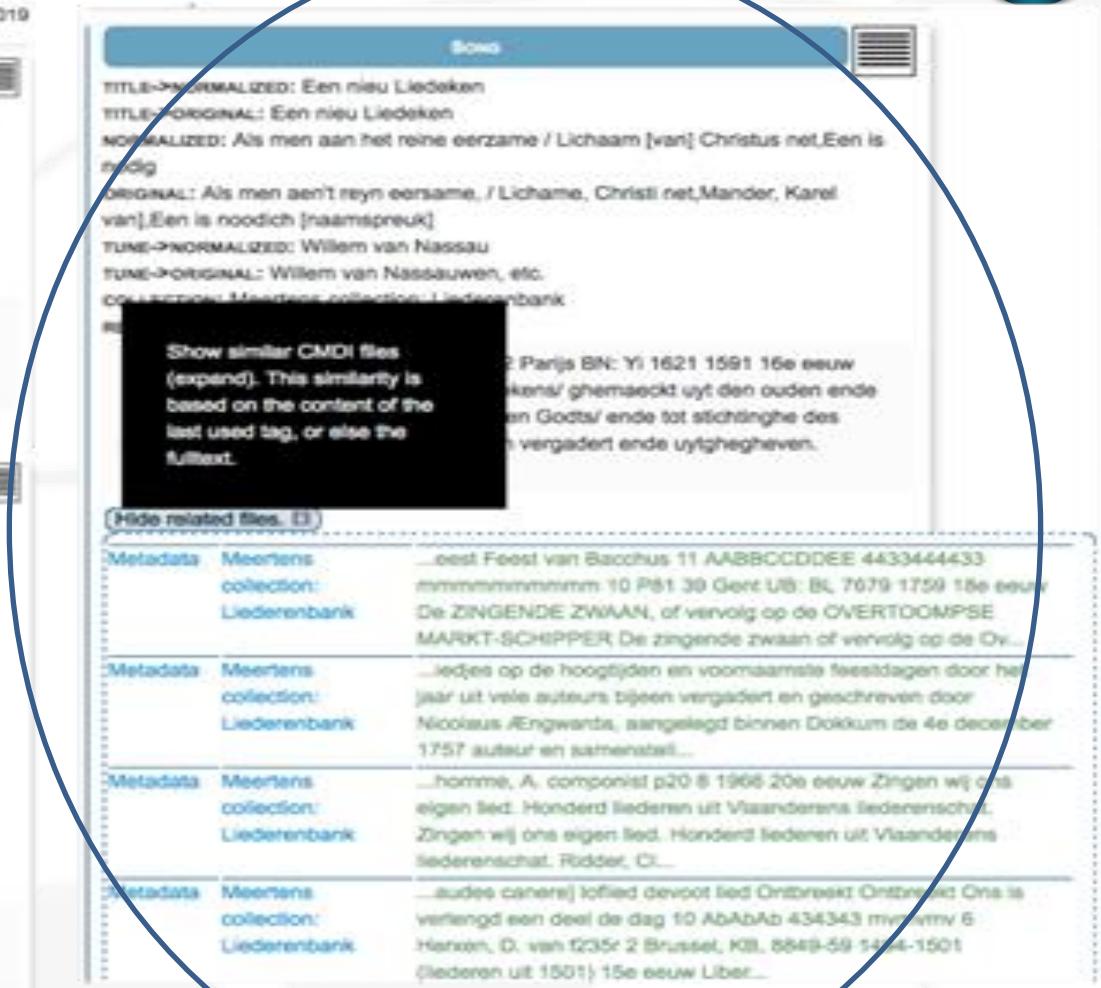


Figure 2b: Recommendation list of related results

Meertens Metadata Search

- Meertens Return Page