CLARIN for Linguists Storing resources in CLARIN

Jan Odijk

LOT Summerschool

Nijmegen, 2014-06-27

Overview

- Why store resource in CLARIN?
- How to store resources in CLARIN
  - What **you** must do
  - What the **CLARIN Centre** must do

Overview

- **Why store resource in CLARIN?**
- How to store resources in CLARIN
  - What **you** must do
  - What the **CLARIN Centre** must do

Why?

- You may benefit from it
  - Existing tools in CLARIN
    - faster production, better quality, more features
    - Search engines, analysis tools, visualisation tools
  - Can be easily combined with other data in CLARIN
- Others may benefit from it
  - Many unexpected uses of your data
  - Now or in the future

Why?

- Openness in Science
  - Usually produced with public money
- Integrity
  - Recently many scandals with faked data
- Verifiability and Replicability of research results
  - Essential for the proper conduct of science
  - More and more journals are requiring it
- Funding Agency requires it (data management plan)
- DANS CLIP on data sharing (in Dutch)

Overview

- Why store resource in CLARIN?
- **How to store resources in CLARIN**
  - What **you** must do
  - What the **CLARIN Centre** must do

How?

- Start early
  - Preferably before you start creating data
- Contact a CLARIN **Type B** centre
  - They can help you
  - Your resource must be stored at a CLARIN centre
- Which CLARIN centre?
  - Check CLARIN-NL portal (http://dev.clarin.nl)
  - CLARIN-NL website Centres page
  - Brief summary

Overview

- Why store resource in CLARIN?
- How to store resources in CLARIN
  - ➢ **What you must do**
  - – What the **CLARIN Centre** must do

What you must do

- Define what your data are / are going to be
- Ensure legal / ethical compliance
  - Permission from subjects to use the data for research
  - Provisions for respecting privacy matters
- Determine Metadata Contents
  - Determine what information should be included in the metadata (resource description) of your resource
  - Collect this information

What you must do

- Determine centre
- Contact them and make arrangements
- Determine CLARIN-recommended standard format(s) for your data
  - Consult with the CLARIN Centre
  - Ask help from the helpdesk helpdesk@clarin.nl
  - E.g. LMF (Lexical Markup Framework for lexicons, cf. Cornetto, DuELME, …)
  -

What you must do

- Metadata must be in CMDI format
- CMDI provides
  - A model for metadata
  - A format for metadata
  - Tools to make metadata
- CMDI Metadata are written in XML
- It does NOT proscribe the contents of the metadata
- Introduction to CMDI cmdi@clarin.eu

What you must do

- CMDI Metadata use a metadata profile
- **Metadata profile**: a combination of
  - Metadata components
  - Metadata elements
- **Metadata Component**: a combination of
  - Metadata components (optional)
  - Metadata elements (optional)
- **Metadata Element**
  - XML element: name, value (of an explicit type), attribute-value pairs

What you must do

- This provides high flexibility
  - YOU can determine the metadata for your resource
  - By defining your own profiles, components, elements
- CMDI helps you with
  - A profile and component editor [login required]
  - A list of commonly used profiles and components
  - A metadata editor: ARBIL

What you must do

- Flexibility requires explicit semantics!
  - The CLARIN infrastructure must `know' what you mean with your metadata elements
  - Otherwise it cannot use faceted browsing in the VLO or the Meertens Metadata Search Engine

What you must do

- Explicit semantics:
  - Each element in the **data and metadata** must have a link to a CLARIN-recognized concept or data category registry
  - Most prominent data category registry in CLARIN is ISOCAT
  - Example Data Category in ISOCAT
  - Example Link to ISOCAT in element definition

What you must do

- Explicit semantics (2):
  - RELCAT alpha version
    - For relations between data categories
  - SCHEMACAT alpha version
    - For describing schemas of resources
  - CLAVAS Vocabulary Service
    - Interface to other data category registries and vocabularies
      - ISO 639-3 Language codes

What you must do

- Attend dedicated tutorials on CMDI and ISOCAT
  - Regularly organized in NL (each 2+ times/year)
- Usually the CLARIN Centre helps you creating the CMDI metadata
- Maximally reuse existing profiles /components
  - It will help you get better metadata
  - You do not have to reinvent the wheel

What you must do

- CLARIN
  - strongly recommends using certain components (e.g. GeneralInfo component) and
  - may require inclusion of certain properties
- Do not forget properties that are ‘obvious to you’, e.g.
  - Language
  - Title
  - Version

What you must do

- Live Version v. exchange/archive version
  - E.g. lexicon
    - in Lexical Markup Framework compatible XML text v.
    - database with indexes for fast search
  - Live version is ideally derived fully automatically from the exchange/archive version
- In close cooperation with the CLARIN Centre

What you must do

- Software
  - Desktop tool
  - Desktop application
  - Web service
    - SOAP / REST , use CLAM if possible
  - Web application
- There must be `metadata' for software as well!
  - Generic profile exists and is being refined
- Consult with the CLARIN Centre

Overview

- Why store resource in CLARIN?
- How to store resources in CLARIN
  - What **you** must do
  - ➢ **What the CLARIN Centre must do**

What the CLARIN Centre must do

- Assist you with your tasks
- Assign Persistent Identifiers (PIDs) to all data and metadata
  - Handle system for assignment and resolution of PIDs
  - example
- Make Metadata harvestable
  - OAI-PMH protocol
    - Open Archives Initiative Protocol for Metadata Harvesting

What the CLARIN Centre must do

- Store the data in the centre's repository
  - LAMUS (the Language Archive) and its documentation online or as PDF
  - EASY (DANS) and its Help and Support Page
- Make data themselves available and accessible in the CLARIN infrastructure

What the CLARIN Centre must do

- Provisions for legal / ethical restrictions
- Long term preservation
  - Data Seal of Approval
- (Minimal) Maintenance

What about existing data?

- Adapt them to meet the CLARIN requirements
  - Data Curation
- CLARIN-NL
  - has financed many data curation projects
  - Has set up Data Curation Service
- CLARIAH (successor project, if financed)
  - Will continue these curation activities

Thanks for your attention!

DO NOT ENTER HERE

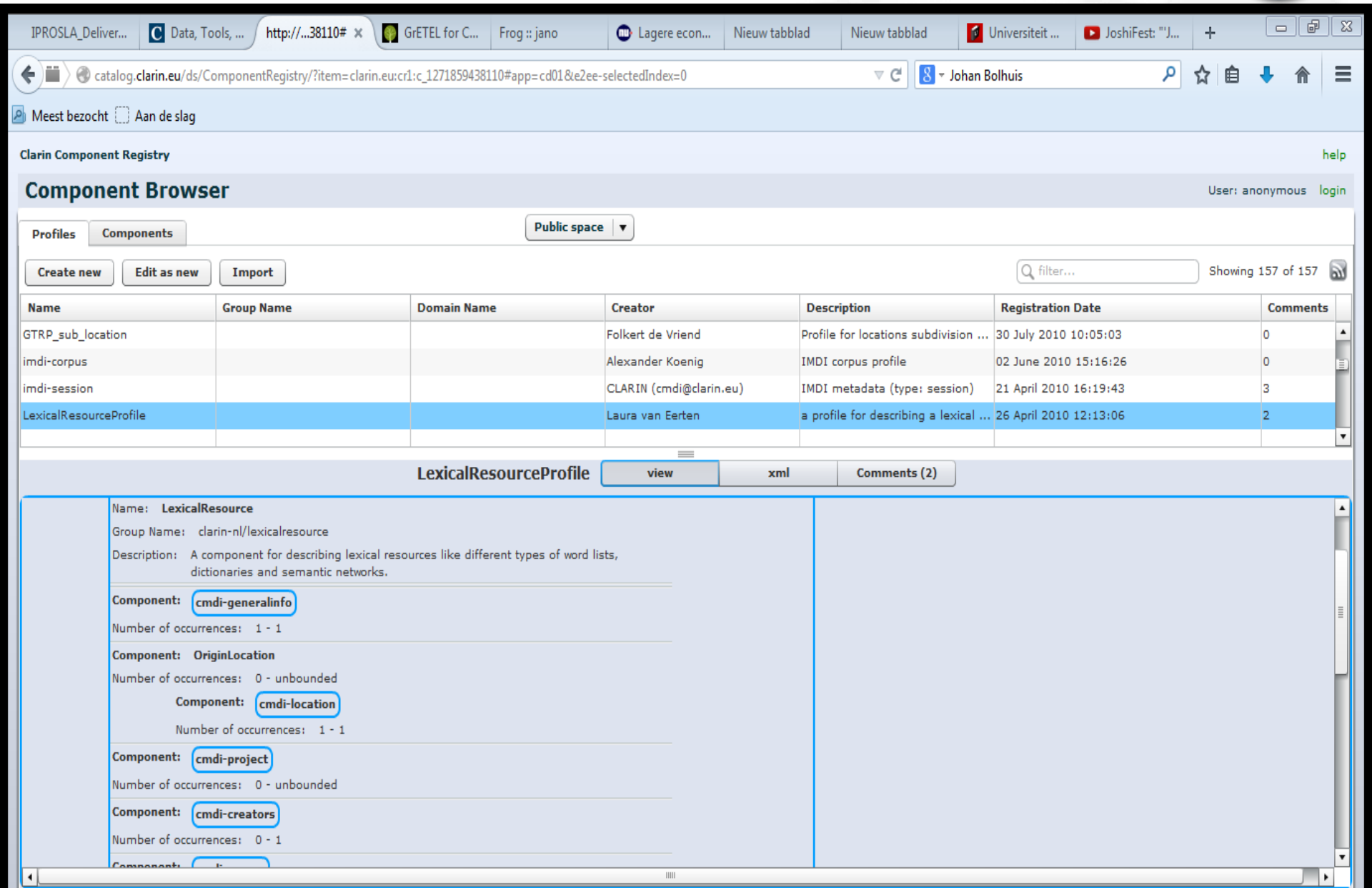
CLARIN-NL B Centres

- **Meertens Institute**: resources relevant for the study of
  - cultural expressions and Language variation within the Dutch language
- **Max Planck Institute for Psycholinguistics (The Language Archive):** resources related to the study of
  - psychological, social and biological foundations of language
- **Huygens Institute**: resources related to the study of
  - history and literature of the Netherlands.
- **Institute for Dutch Lexicology (INL)**
  - relevant to the lexicological study of the Dutch language
- **Data Archiving and Networked Centres (DANS)**
  - digital research data generally

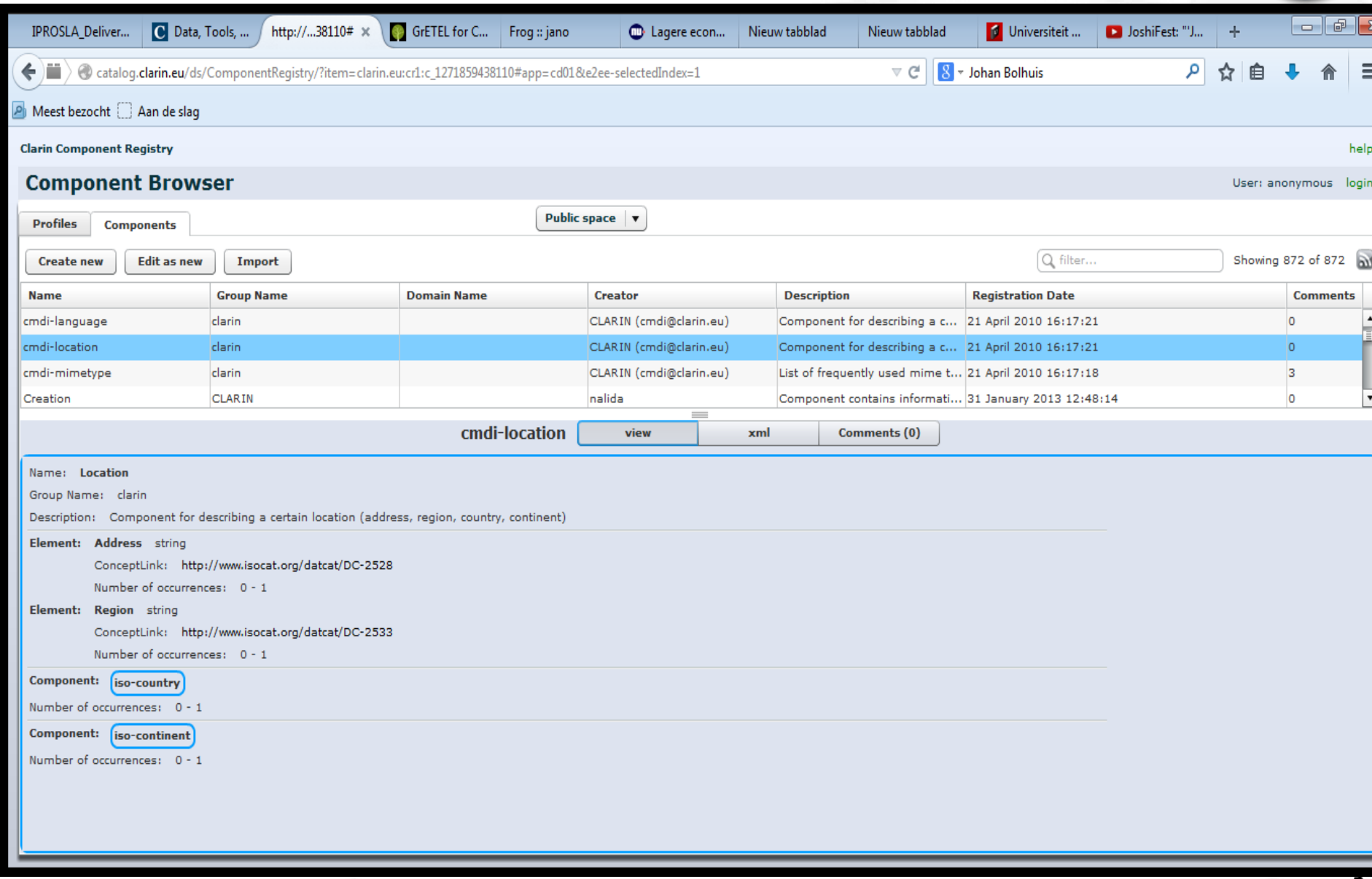
CLARIN-NL D Centres

- **Koninklijke Bibliotheek (National Library)**:
  - Digital books, articles, newspapers
  - Includes DBNL (Digital Library for Dutch Literature)
  - (will be available in the VLO soon)
- **Nederlands Instituut voor Beeld & Geluid (NIBG, Netherlands Institute for Sound and Vision)**
  - Audio-visual data (esp. TV and radio programmes)
  - NIBG data via the VLO
- **Utrecht University Library (UBU)**
  - Digital books, articles
  - UBU data via the VLO

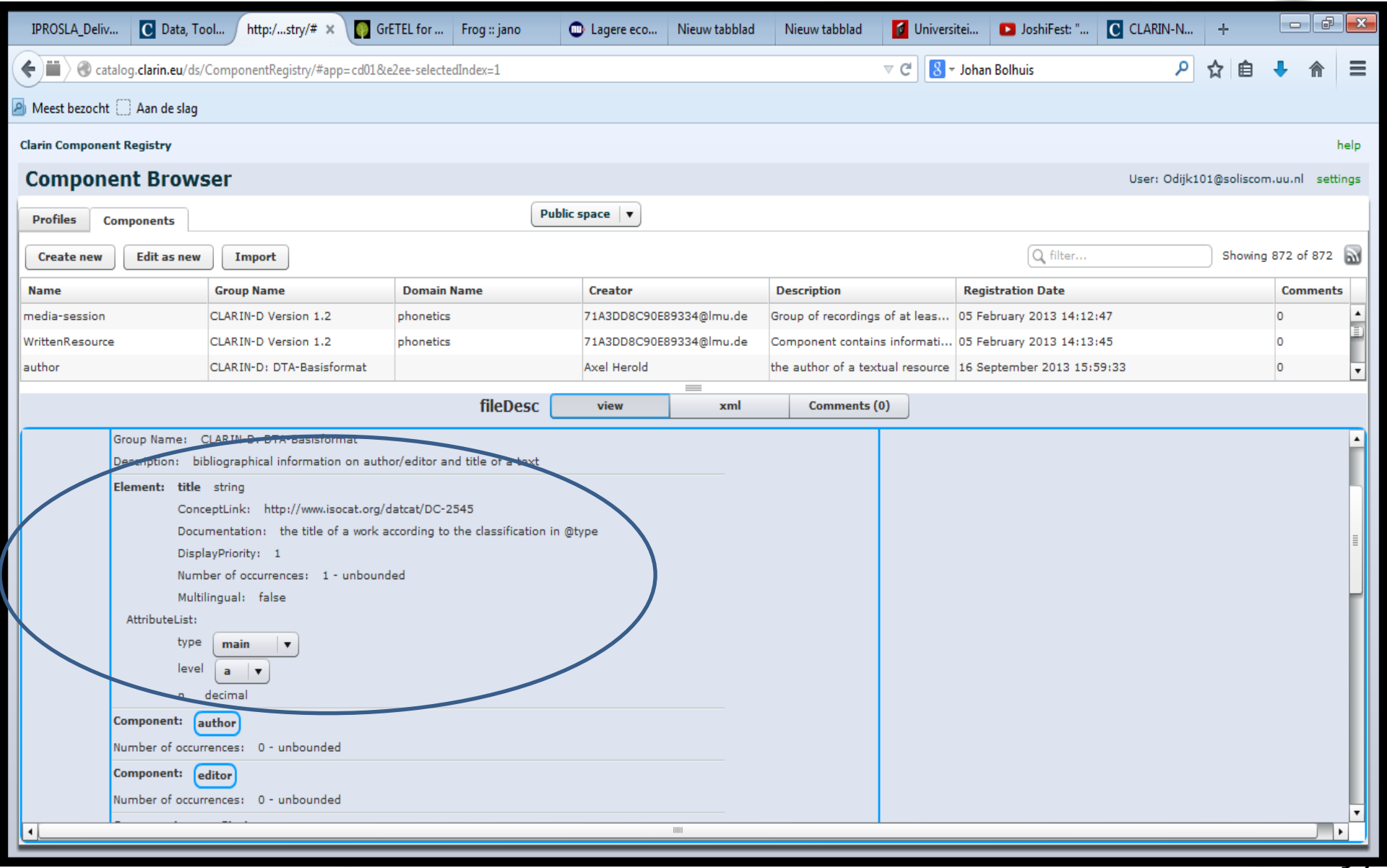
CMDI Profile: Example



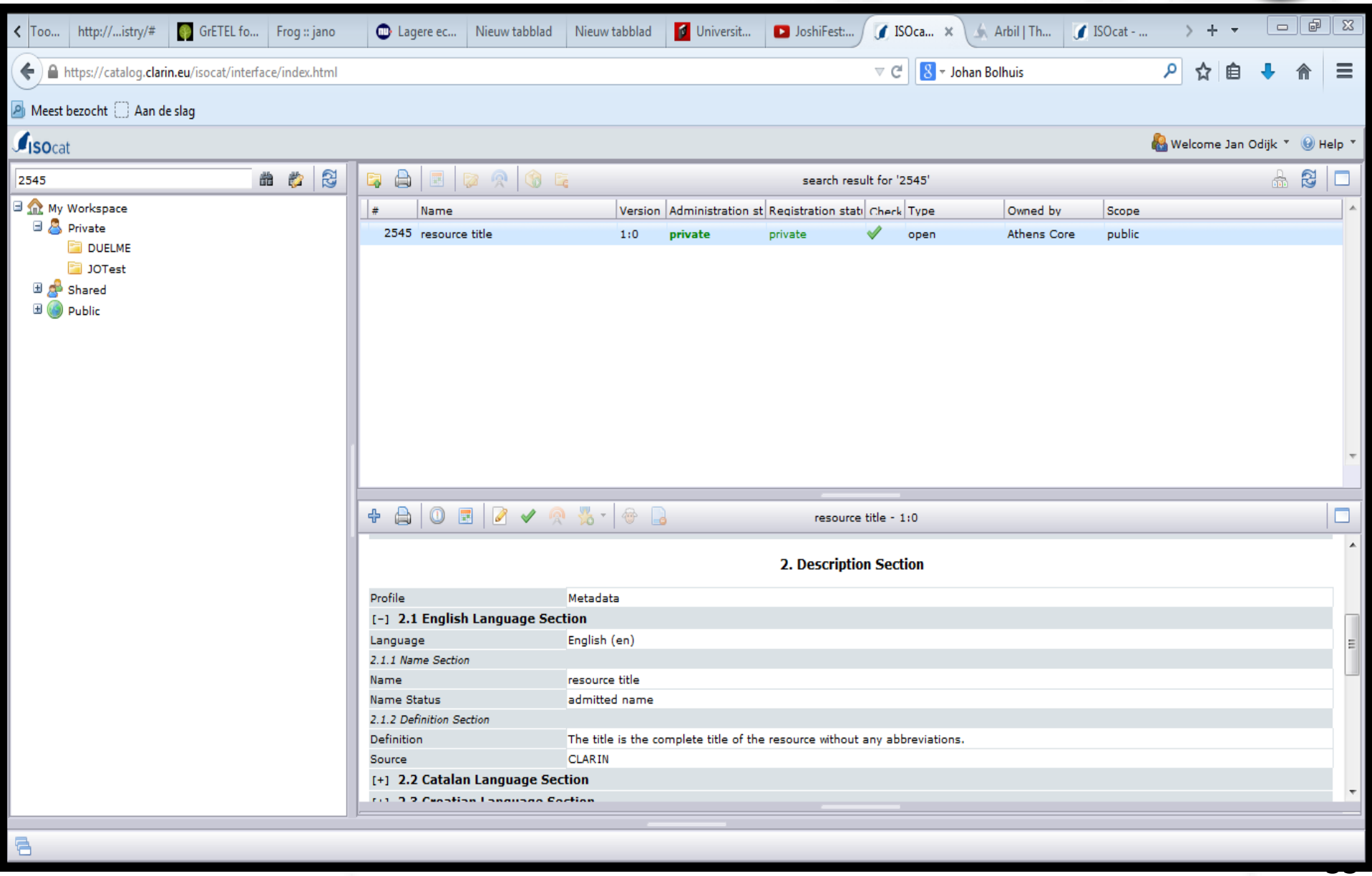
CMDI Component: Examp



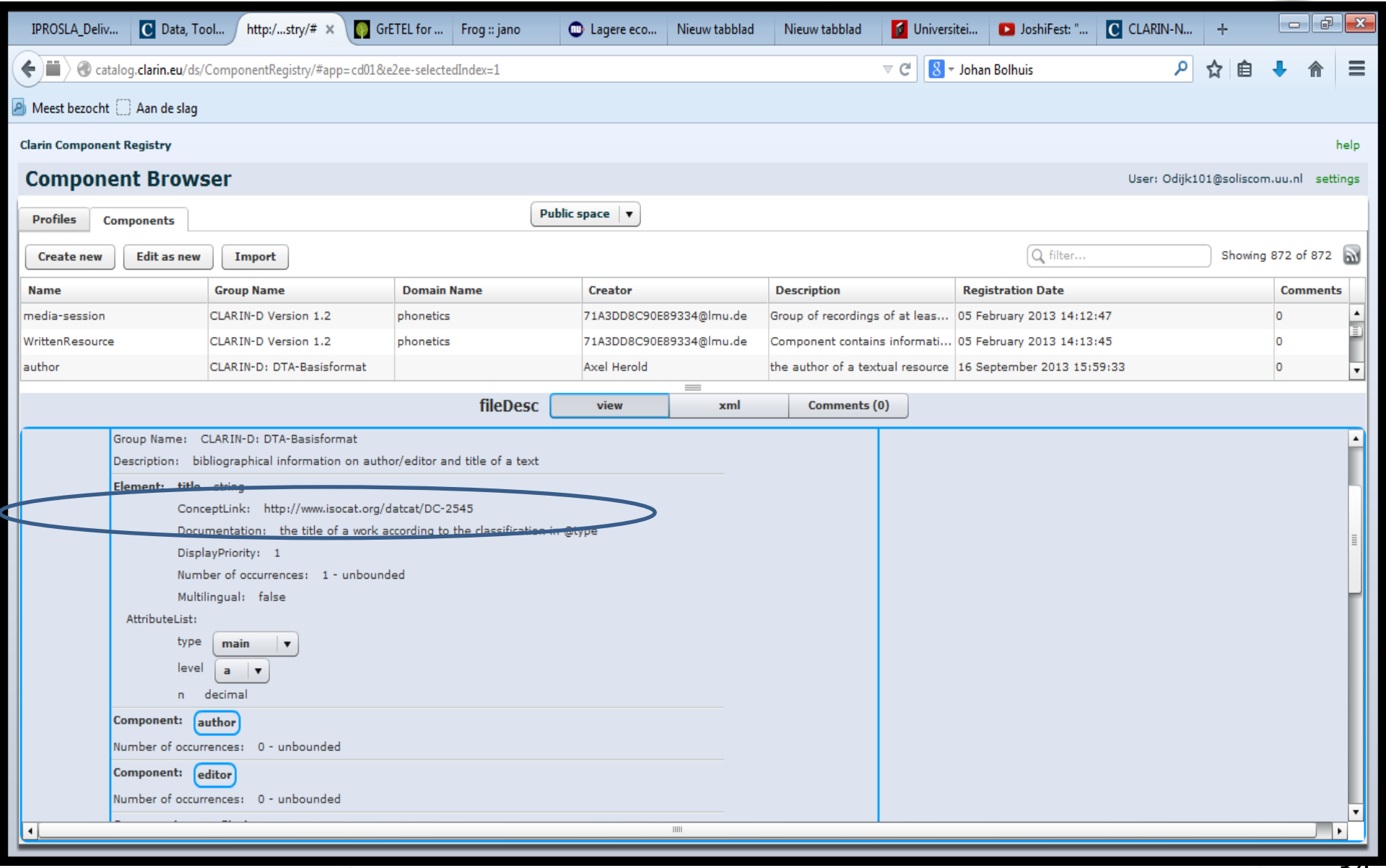
CMDI Element: Exampl



ISOCAT DC


https://catalog.clarin.eu/isocat/interface/index.html
Johan Bolhuis
Meest bezocht
Aan de slag
ISOcat
Welcome Jan Odijk
Help
2545
My Workspace
Private
DUELME
JOTest
Shared
Public
search result for '2545'
Name Version Administration st Registration stati Check Type Owned by Scope
2545 resource title 1:0 private private open Athens Core public
resource title - 1:0
2. Description Section
Profile Metadata
[-] 2.1 English Language Section
Language English (en)
2.1.1 Name Section
Name resource title
Name Status admitted name
2.1.2 Definition Section
Definition The title is the complete title of the resource without any abbreviations.
Source CLARIN
[+] 2.2 Catalan Language Section

LINK TO ISOCAT DC



LOGIN

- CLARIN attempts to maximize open and free access to resources
  - with as little restrictions as possible
  - no login unless it cannot be avoided
- Sometimes, a login (Authentication and Authorisation, AAI) is required, e.g.
  - Because there are legal and/or ethical restrictions on the data
  - To identify you and assign you your own workspace / data
  - To enter with your own personal settings

LOGIN

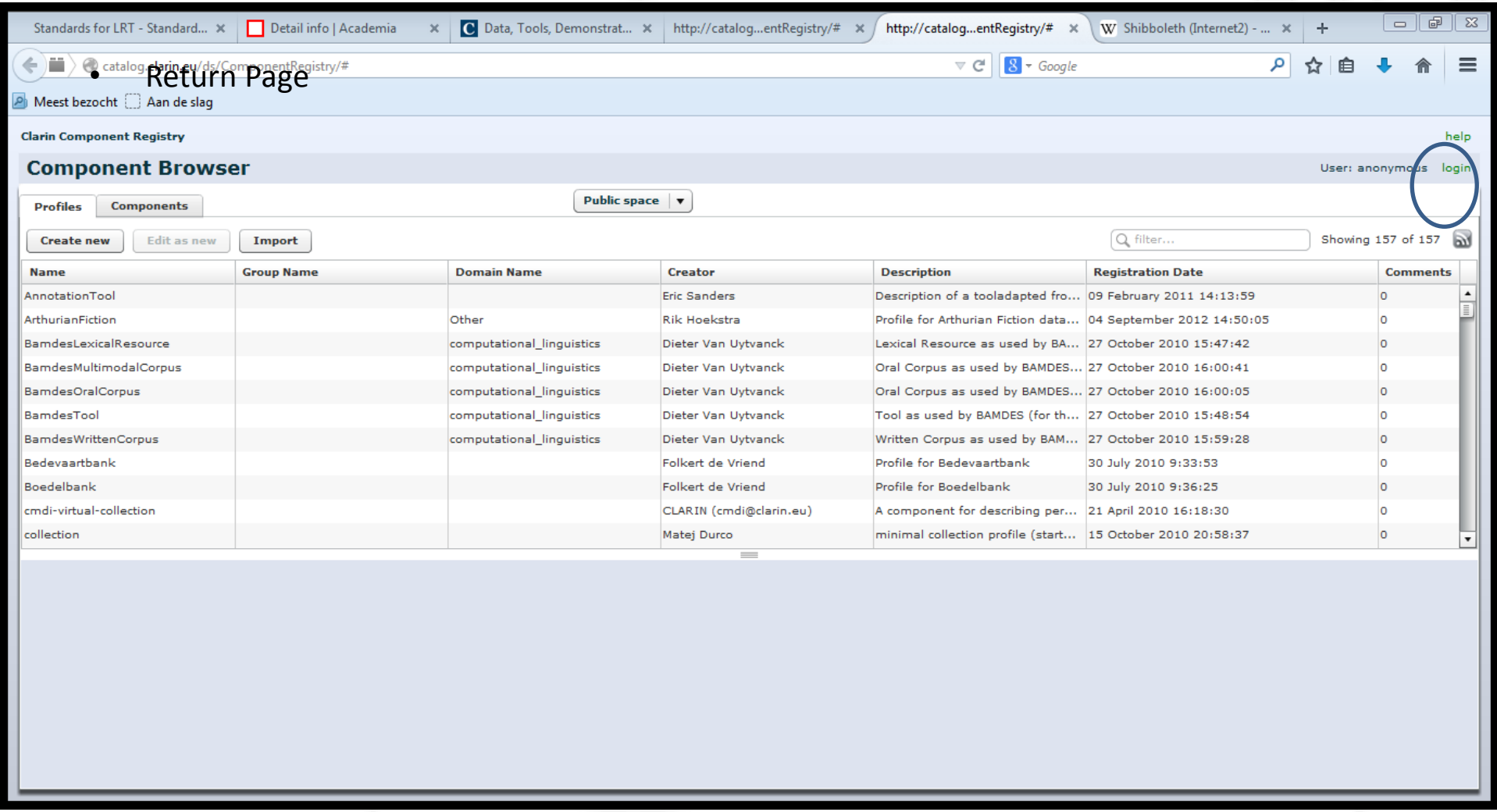
- CLARIN is a distributed infrastructure
  - How can we avoid that you have to login again and again?
  - How can we avoid that you have to remember many user names and passwords?
  - How can we avoid that CLARIN has to securely store user names, passwords and possibly other privacy-sensitive information?

LOGIN

- The answer: Shibboleth
  - When you log in, you are directed to a login with your own institute
  - You then log in with you institute's user name and password
  - The institute server then confirms that you are a trusted person, and you can enter this part of the CLARIN infrastructure
    - it does **not** pass on any sensitive information such as your user name or password
  - If you now go to another part of the CLARIN infrastructure that requires login, it `knows' that you are already logged in, so you do not have to do this again (Single Sign On, SSO)

LOGIN

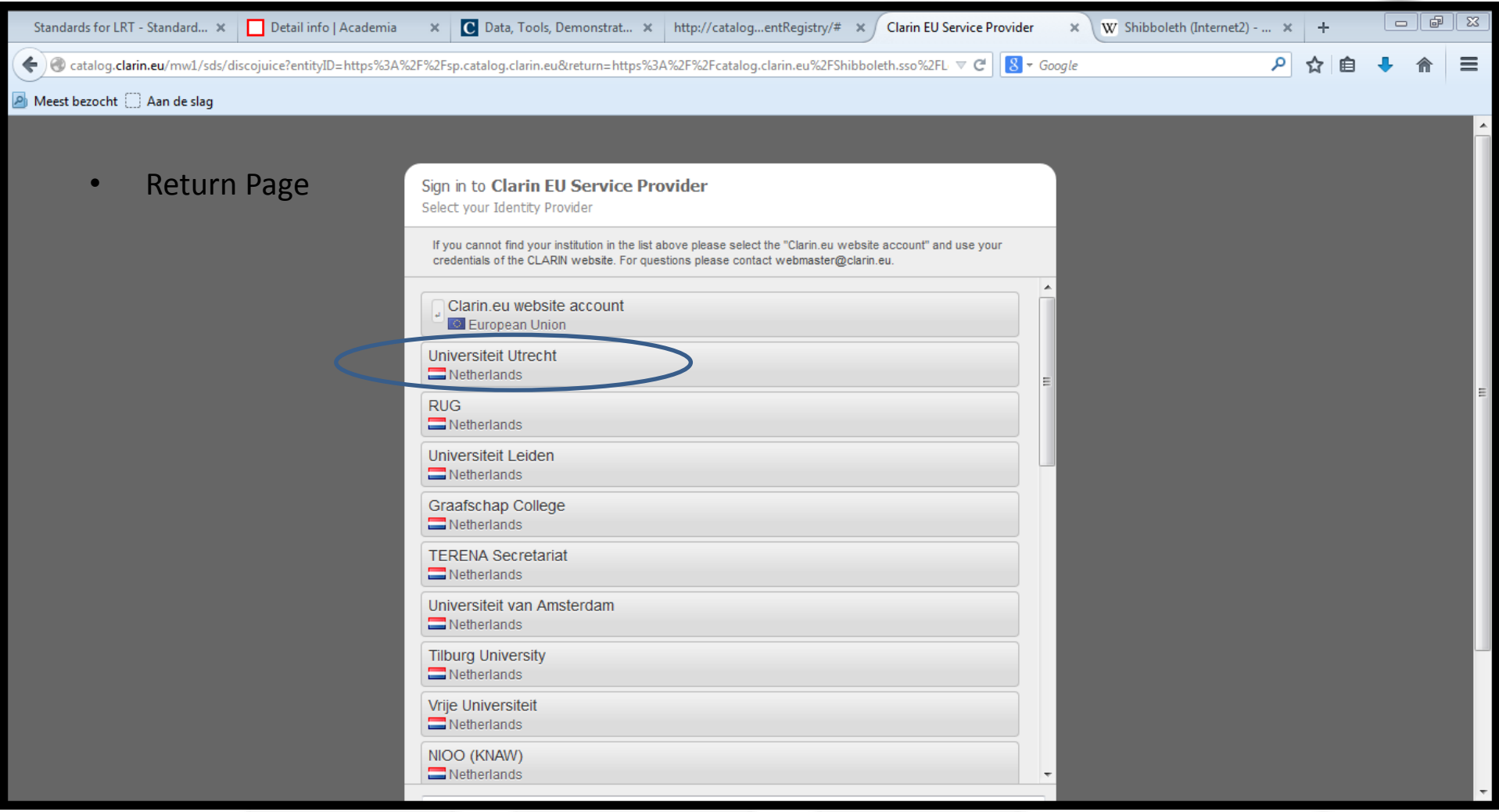
- Return Page



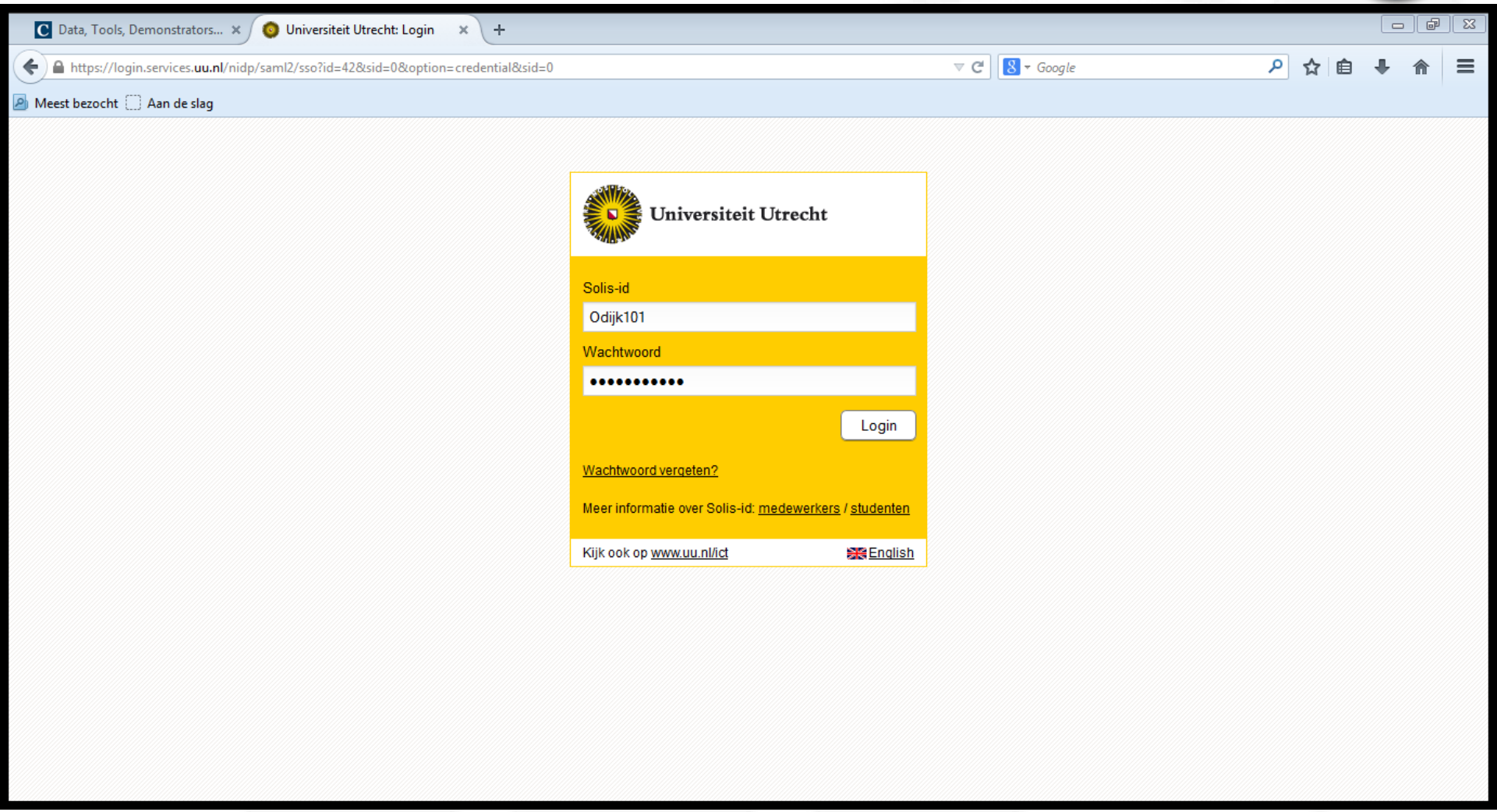
Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
AnnotationTool			Eric Sanders	Description of a tooladapted fro...	09 February 2011 14:13:59	0
ArthurianFiction		Other	Rik Hoekstra	Profile for Arthurian Fiction data...	04 September 2012 14:50:05	0
BamdesLexicalResource		computational_linguistics	Dieter Van Uytvanck	Lexical Resource as used by BA...	27 October 2010 15:47:42	0
BamdesMultimodalCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:41	0
BamdesOralCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:05	0
BamdesTool		computational_linguistics	Dieter Van Uytvanck	Tool as used by BAMDES (for th...	27 October 2010 15:48:54	0
BamdesWrittenCorpus		computational_linguistics	Dieter Van Uytvanck	Written Corpus as used by BAM...	27 October 2010 15:59:28	0
Bedevaartbank			Folkert de Vriend	Profile for Bedevaartbank	30 July 2010 9:33:53	0
Boedelbank			Folkert de Vriend	Profile for Boedelbank	30 July 2010 9:36:25	0
cmdi-virtual-collection			CLARIN (cmdi@clarin.eu)	A component for describing per...	21 April 2010 16:18:30	0
collection			Matej Durco	minimal collection profile (start...	15 October 2010 20:58:37	0

LOGIN

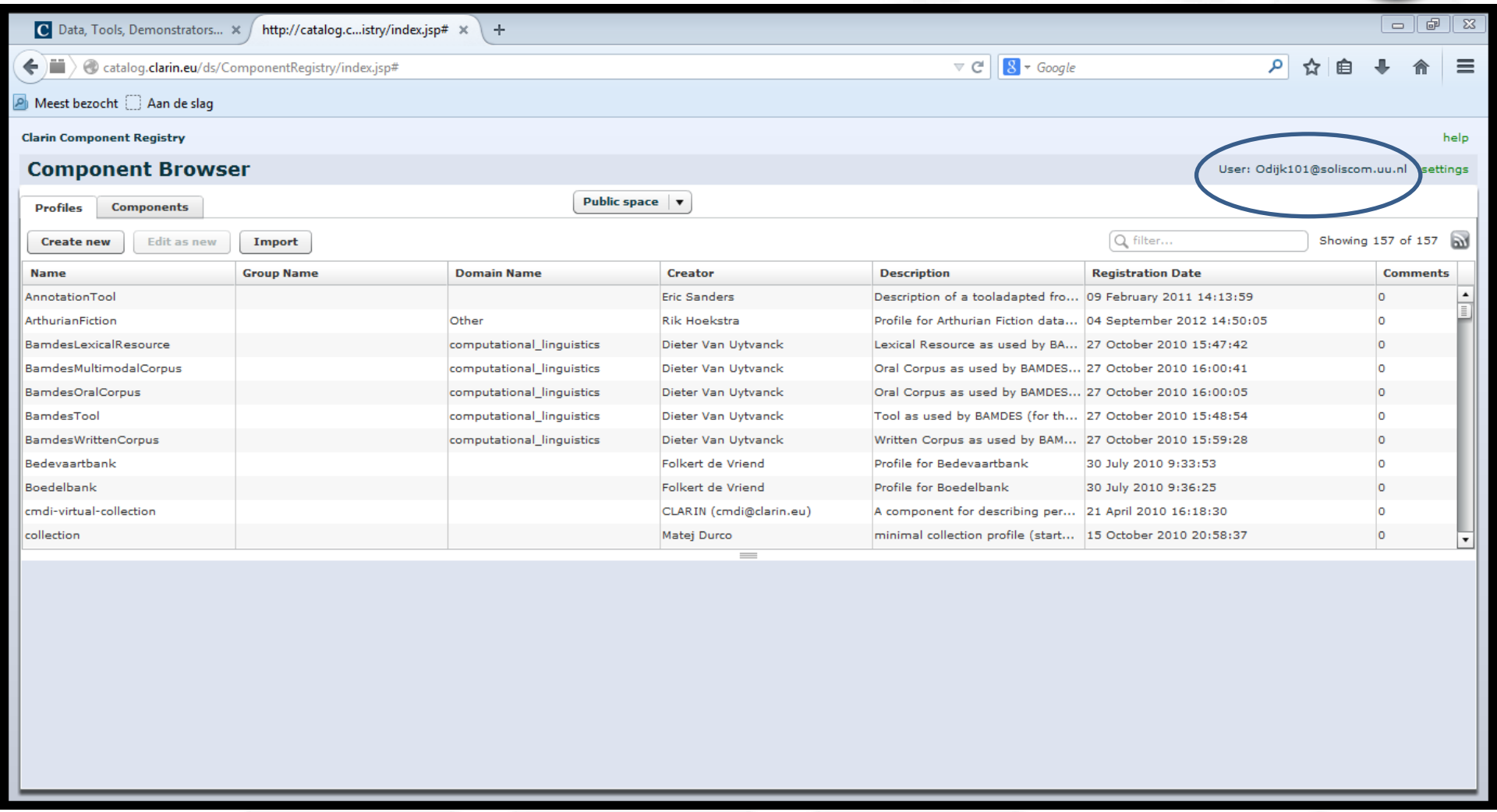
- Return Page



LOGIN

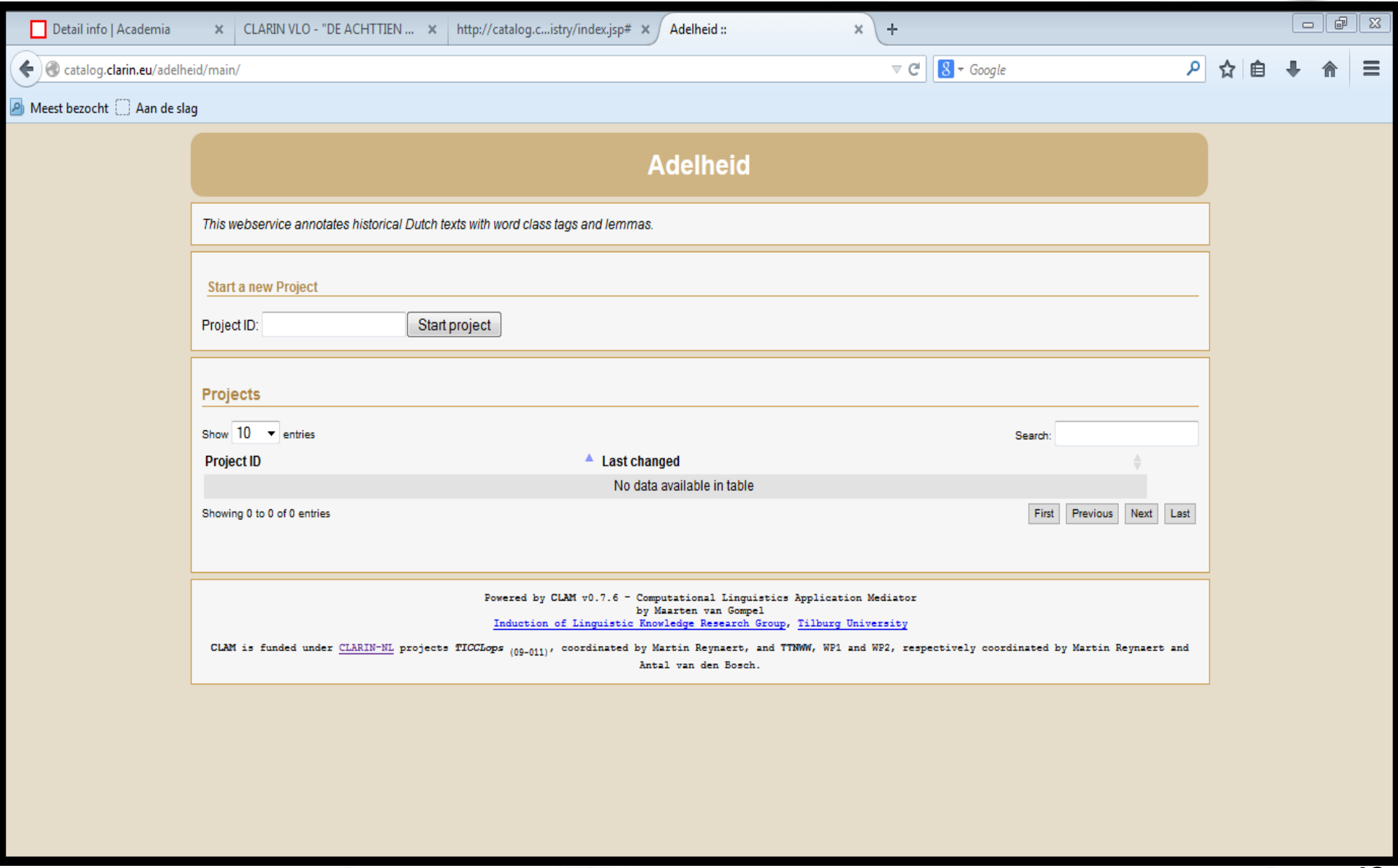

Data, Tools, Demonstrators...
Universiteit Utrecht: Login
https://login.services.uu.nl/nidp/saml2/sso?id=42&sid=0&option=credential&sid=0
Google
Meest bezocht
Aan de slag
Universiteit Utrecht
Solis-id
Odijk101
Wachtwoord
Login
Wachtwoord vergeten?
Meer informatie over Solis-id: medewerkers / studenten
Kijk ook op www.uu.nl/ict
English

LOGIN



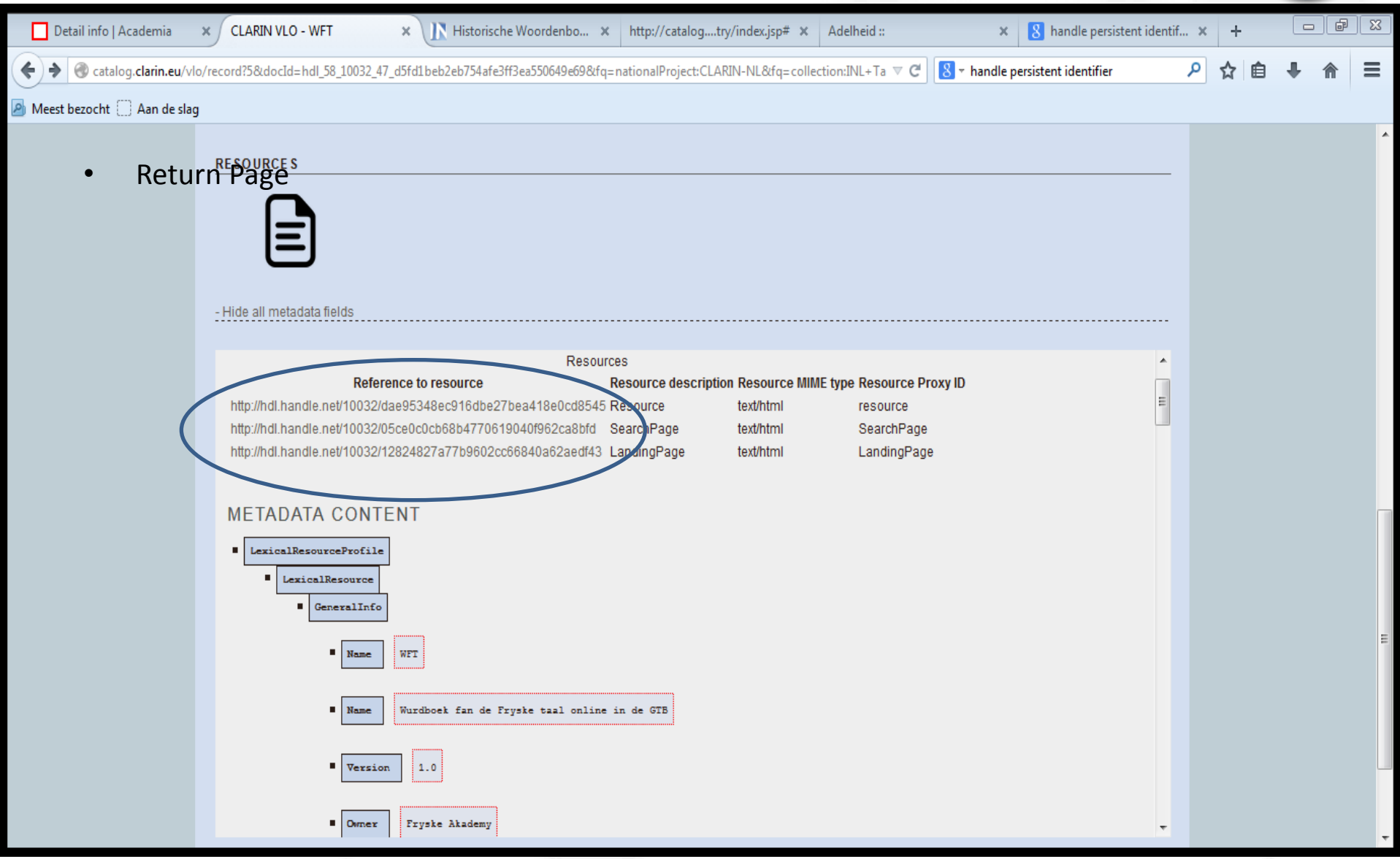
Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
AnnotationTool			Eric Sanders	Description of a tooladapted fro...	09 February 2011 14:13:59	0
ArthurianFiction		Other	Rik Hoekstra	Profile for Arthurian Fiction data...	04 September 2012 14:50:05	0
BamdesLexicalResource		computational_linguistics	Dieter Van Uytvanck	Lexical Resource as used by BA...	27 October 2010 15:47:42	0
BamdesMultimodalCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:41	0
BamdesOralCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:05	0
BamdesTool		computational_linguistics	Dieter Van Uytvanck	Tool as used by BAMDES (for th...	27 October 2010 15:48:54	0
BamdesWrittenCorpus		computational_linguistics	Dieter Van Uytvanck	Written Corpus as used by BAM...	27 October 2010 15:59:28	0
Bedevaartbank			Folkert de Vriend	Profile for Bedevaartbank	30 July 2010 9:33:53	0
Boedelbank			Folkert de Vriend	Profile for Boedelbank	30 July 2010 9:36:25	0
cmdi-virtual-collection			CLARIN (cmdi@clarin.eu)	A component for describing per...	21 April 2010 16:18:30	0
collection			Matej Durco	minimal collection profile (start...	15 October 2010 20:58:37	0

LOGIN


Detail info | Academia
CLARIN VLO - "DE ACHTTIEN ...
http://catalog.c...istry/index.jsp#
Adelheid ::
catalog.clarin.eu/adelheid/main/
Google
Meest bezocht
Aan de slag
Adelheid
This webservice annotates historical Dutch texts with word class tags and lemmas.
Start a new Project
Project ID:
Start project
Projects
Show 10 entries
Search:
Project ID
Last changed
No data available in table
Showing 0 to 0 of 0 entries
First
Previous
Next
Last
Powered by CLAM v0.7.6 - Computational Linguistics Application Mediator
by Maarten van Gompel
Induction of Linguistic Knowledge Research Group, Tilburg University
CLAM is funded under CLARIN-NL projects TICCLops (09-011), coordinated by Martin Reynaert, and TTNWW, WP1 and WP2, respectively coordinated by Martin Reynaert and Antal van den Bosch.

PIDs

- Return Page



LOGIN

- Return Page