The CLARIN infrastructure in the Netherlands: What is it and how can you use it?

Jan Odijk
UiL-OTS, Utrecht University
j.odijk@uu.nl

1 Introduction

In this paper I will describe what the CLARIN infrastructure is and how it can be used, with a focus on the Netherlands part of the CLARIN infrastructure. I aim to explain how a humanities researcher can use the CLARIN infrastructure.[1] A separate paper describes what had to be done behind the scenes to make this work Odijk (2014f).

CLARIN was prepared by the CLARIN preparatory project (CLARIN-PP, 2008-2011), funded by the European Commission and coordinated by Utrecht University. Since February 2012 CLARIN is coordinated by CLARIN ERIC, hosted by the Netherlands. An *ERIC (European Research Infrastructure Consortium)* is a legal entity at the European level specifically set up for European research infrastructures. Apart from the Netherlands, the other CLARIN ERIC members currently are Austria, Bulgaria, the Czech Republic, Denmark, the Dutch Language Union, Estonia, Germany, and Poland, with Norway as an observer. Many other countries are at the verge of joining CLARIN ERIC, e.g. Sweden and Finland, and the ERIC is expected to grow larger in the coming years. Each ERIC member commits to paying the ERIC yearly fee and to contributing to the CLARIN infrastructure by setting up national projects to this end. Most of the work reported on in this paper has been carried out in the Netherlands national CLARIN project (called *CLARIN-NL*), which runs from 2009 through 2014.

I will first describe what CLARIN is and what it is intended for (section 2). It will include a description of the major functionality CLARIN aims to offer to researchers: finding data and software (described in section 4), applying software to data (described in section 5), storing data and software in the CLARIN infrastructure (described in section 6, and all of that via single portal (described in section 3). I end with concluding remarks (section 7).

2 The CLARIN Infrastructure

The CLARIN infrastructure (from now on simply *CLARIN*) is a **research infrastructure** for **humanities researchers** who work with **digital language resources**. I will explain each of the bold-faced terms.

[1] There is a series of presentations covering the major contents of this paper: (Odijk, 2014b, Odijk, 2014c, Odijk, 2014d, Odijk, 2014e, Odijk, 2014a).

Infrastructure refers to (usually large-scale) basic physical and organizational resources, structures and services needed for the operation of a society or enterprise.[2] Familiar examples are the railway network (figure 1), the road network, the electricity network, but also (on a smaller scale) Eduroam[3], which provides world-wide wireless internet facilities through organisations for higher and further education.

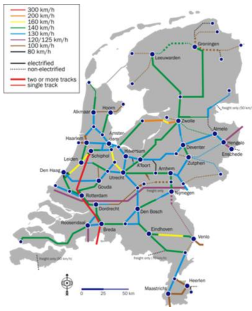


Figure 1: Dutch Railway Network

A **research infrastructure** is an infrastructure intended for carrying out research: facilities, resources and related services used by the scientific community to conduct top-level research. Famous examples are the Chile large telescope (figure 2) and the CERN Large Hadron Collider (figure 3).

Figure 2: Chile Large Telescope

Humanities researchers include linguists, historians (including art historians), literary scholars, philosophers, religion scholars, and others, as well as political science researchers, who are usually considered part of the social sciences.[4]

[2] This description is an adaptation of the description from (English) Wikipedia `http://en.wikipedia.org/wiki/Infrastructure`.

[3] I will provide hyperlinks in the text but usually not show the URL. People reading this paper electronically can directly click on such links. People who read this article on paper do not want to copy the URLs by hand anyway, so they will turn to the electronic version if they want to follow a link.

[4] CLARIN at the European scale is intended for the humanities *and* the social sciences, but the Nether-

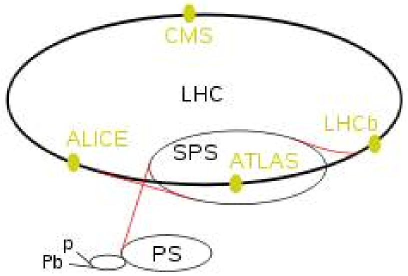


Figure 3: Large Hadron Collider

Digital language resources includes both data and software. It includes a wide spectrum of digital data types:

- Data in natural language (texts, lexicons, grammars, etc.)
- Databases about natural language (typological databases, dialect databases, lexical databases, etc.)
- Audio-visual data containing (written, spoken, signed) language (e.g. pictures of manuscripts, audiovisual data for language description, description of sign language, interviews, radio and tv programmes, etc.)

As for software, digital language resources include software dedicated to browse and search in digital language data (e.g. software to search in a linguistically annotated text corpus), as well as software to analyze, enrich, process, and visualize digital language data, (e.g., a parser, which enriches each sentence in a text corpus with a syntactic structure). I will often use the short term *resource* instead of *digital language resource*.
CLARIN is intended for **language** in various functions, including:

- As an object of inquiry
- As a carrier of cultural content
- As a means of communication
- As a component of identity

Though the creation of data for research certainly is part of creating a research infrastructure, CLARIN has **not** created any new data.[5] It has mainly adapted existing data and software, and it has created new easy and user-friendly software for searching, analysing and visualising data.

CLARIN is not one big physical installation on a single location such as the CERN Large Hadron Collider or the Chile Large Telescope. On the contrary,

lands has focused on the humanities.

[5] This is certainly true for the Netherlands, and to my knowledge also for most other countries involved in CLARIN.

- CLARIN is a **distributed** infrastructure: it has been implemented as a network of **CLARIN centres**. The Netherlands has several such centres. These will be discussed in more detail in section 6.2.
- CLARIN is a **virtual** infrastructure: it provides services electronically (via the internet). Every user can use CLARIN from any location where he/she has access to internet.[6]

The CLARIN infrastructure is still under construction, is highly incomplete, is fragile in some respects, but many parts of it can already be used.

The CLARIN infrastructure offers services so that a researcher

- Can find all data and software relevant for the research
- Can apply the software to the data without any technical background or ad-hoc adaptations
- Can store data and and software resulting from the research

- via one portal

I will discuss each of these aspects in the sections to follow: the portal in section 3, finding data and software in section 4, applying software to the data in section 5, and storing data and software in the CLARIN infrastructure in section 6.

3 Portal

It is convenient for users if they do not have to remember a lot of URLs or other identifiers to get access to the functionality offered by CLARIN. For this reason, a portal has been set up for CLARIN. The idea is that from this portal all functionality offered by CLARIN can be accessed.

The Europe-wide CLARIN portal can be found via the CLARIN website, top menu item *Portal*, or directly.

The CLARIN portal gives access to the Virtual Language Observatory (see section 4.1), featured resources, showcases, general information on CLARIN, CLARIN-related blogs, instructions on how to deposit your resources, and it offers the opportunity to search through multiple corpora with one query.

In addition to the Europe-wide portal, also national CLARIN portals are being created.[7] These also will make it possible to access all CLARIN functionality but will put special emphasis on data and software created nationally. The national CLARIN portal for the Netherlands is currently under construction and can temporarily be accessed via the URL `dev.clarin.nl` and later via `www.clarin.nl`.[8]

[6] Though CLARIN also makes available software that operates locally on a single computer. This is necessary in some cases where internet access is absent or limited.

[7] It is not a problem that there are multiple portals, which each put the focus on different aspects of the CLARIN infrastructure. However, it is essential that all functionality in CLARIN can be reached from each portal. And at least one portal, the CLARIN ERIC portal, should contain links to all other portals.

[8] While the portal is under construction, a complete list of the results (data, web applications, services) of the CLARIN-NL project and links to them can be obtained via `http://www.clarin.nl/node/404`.

The Dutch national portal offers an introductory page, an overview of Dutch CLARIN centres, a selection of tools to find relevant resources through their metadata and to search in data themselves, an inventory of tools and services with faceted search on facets such as resource type, relevant scientific disciplines, tool functionality, and others. For example, if you are interested in *syntax*, select that value for the facet *research discipline*; if, within syntax, you are more specifically interested in *parsing*, you can select this value for the facet *toolTask*: one then ends up with descriptions of the INPOLDER parser for 13th century Dutch and for the *Alpino* parser for Modern Dutch that is offered via *TTNWWW*.These descriptions also contain links to the actual services, their documentation and demonstration scenarios. See figure 4.

The portal also offers a section called *CLARIN recipes* to get concrete instructions on how to carry out frequent actions, and it offers an opportunity to ask colleagues for advice.

4 Finding data and software

An essential function offered by CLARIN is the possibility to find resources (data and software) that might be relevant to your research. That is in itself not a trivial task, but it is especially difficult because of the distributed character of the CLARIN infrastructure. How can one find data and software that are distributed over multiple CLARIN centres? Of course, access is possible via the internet, but, as is well-known, web pages and URLs regularly change or even disappear over time: how can it be guaranteed that a link to data is there still tomorrow? Searching via Google will not work, because even if it finds all relevant results, it will also find too many irrelevant search results, and it will not be easy and a lot of work to select the relevant ones.

CLARIN offers this functionality as follows. First, it offers for all resources descriptions of the resources (also known as *metadata*). Such *resource descriptions*[9] are made in the *CMDI* format. *CMDI* stands for *Component-based Metadata Infrastructure* and it offers a flexible format for representing descriptions of resources. CMDI prescribes the format of the resource descriptions but not their contents: these are determined by the data provider. I will go deeper into CMDI in section 6.

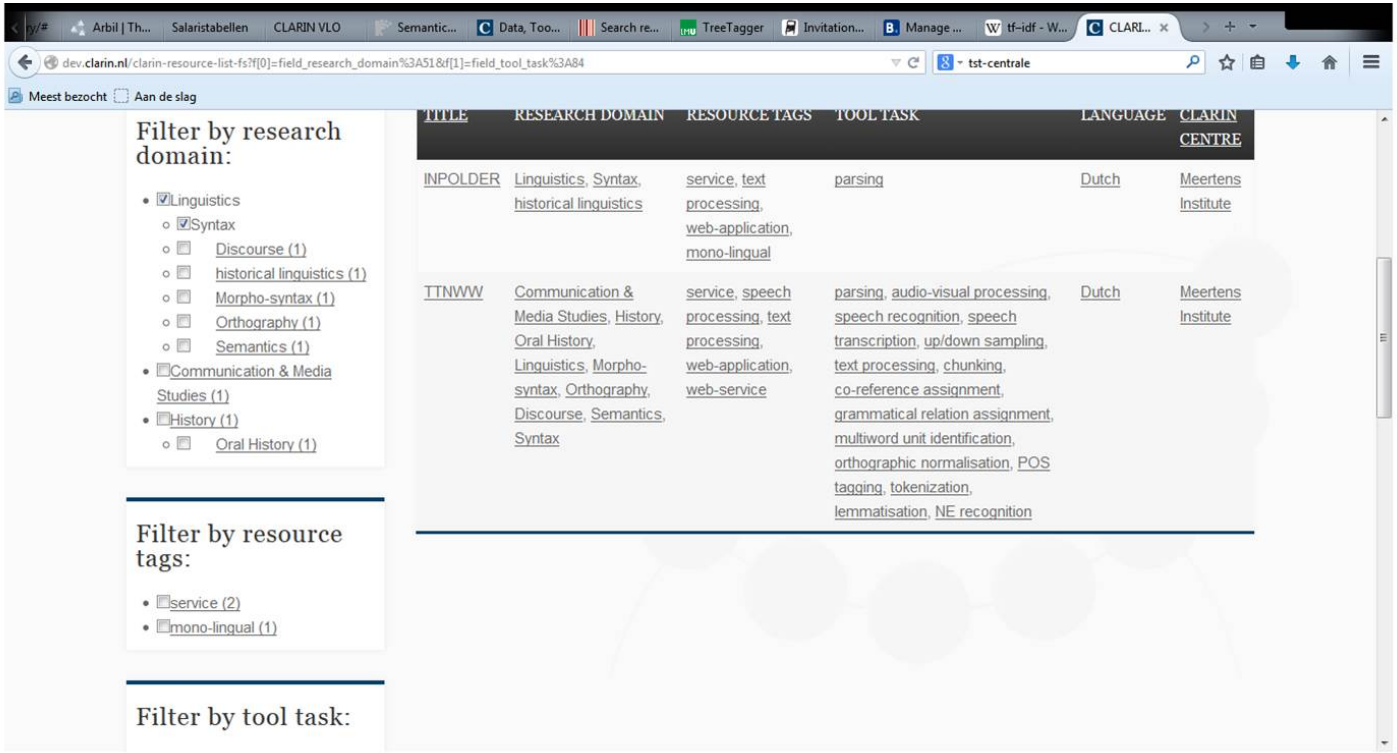
Second, the resources and their CMDI-descriptions are stored on servers of CLARIN Centres. The CMDI-descriptions are made available to the outside world via a specific protocol, the *OAI-PMH* protocol (*Open Archives Initiative - Protocol for Metadata Harvesting*).

Third, all resource descriptions and all resources will be referred to via *persistent identifiers (PIDs)*, i.e identifiers that are guaranteed to exist and correctly refer persistently.

Fourth, CLARIN offers browsers and search engines to browse and search for resources via their CMDI resource descriptions. Such browsers and search engines operate on a database of CMDI resource descriptions located on a server of a specific CLARIN centre. This database is filled and regularly updated[10] by 'metadata harvesting', i.e. an automatic process of collecting all resource descriptions made available by the various CLARIN centres (using the OAI-PMH protocol) and storing them in a single database.

[9] In this paper I will systematically use the term *resource description* and avoid the term *metadata* for the reasons sketched in Odijk and van Hessen (2011:100).

[10] See `http://www.clarin.eu/faq/when-metadata-vlo-harvested` for the update schedule for one such search engine, the *Virtual Language Observatory.*



Filter by research domain:

- ☑Linguistics
  - ☑Syntax
  - ☐ Discourse (1)
  - ☐ historical linguistics (1)
  - ☐ Morpho-syntax (1)
  - ☐ Orthography (1)
  - ☐ Semantics (1)
- ☐Communication & Media Studies (1)
- ☐History (1)
  - ☐ Oral History (1)

Filter by resource tags:

- ☐service (2)
- ☐mono-lingual (1)

Filter by tool task:

TITLE	RESEARCH DOMAIN	RESOURCE TAGS	TOOL TASK	LANGUAGE	CLARIN CENTRE
INPOLDER	Linguistics, Syntax, historical linguistics	service, text processing, web-application, mono-lingual	parsing	Dutch	Meertens Institute
TTNWW	Communication & Media Studies, History, Oral History, Linguistics, Morpho-syntax, Orthography, Discourse, Semantics, Syntax	service, speech processing, text processing, web-application, web-service	parsing, audio-visual processing, speech recognition, speech transcription, up/down sampling, text processing, chunking, co-reference assignment, grammatical relation assignment, multiword unit identification, orthographic normalisation, POS tagging, tokenization, lemmatisation, NE recognition	Dutch	Meertens Institute

Figure 4: Selection of serviced via faceted browsing in the Dutch portal

Currently, CLARIN offers two browsers and search engines to search for resources via their resource descriptions, viz. the *Virtual Language Observatory (VLO)*, which will be discussed in section 4.1, and the *Meertens CLARIN Metadata Search Engine*, which will be discussed in section 4.2.

Which resources can you find in the the CLARIN infrastructure? There are several. First, there are the data and software of the CLARIN centres themselves (e.g. the Corpus Gysseling and search engine at INL). Second, there are the data and software hosted by a centre but originating from a researcher from another research organisation (e.g. the FESLI data and search engine at Meertens). Third, there are CLARIN Centres of a special type (called *Data Providers* or Type D CLARIN Centres[11]), which distribute data independently of (and long before) CLARIN, but have made provisions to give access to the data that are relevant to humanities researchers in a CLARIN-compatible manner (via CMDI resource descriptions). These CLARIN Centres include organisations that, by their very mission, make available large amounts of data, currently:

Koninklijke Bibliotheek (KB) [12] Digital books, articles, newspapers. Includes the DBNL (Digital Library for Dutch Literature)

Nederlands Instituut voor Beeld & Geluid (NIBG) [13] Audio-visual data (esp. TV and radio programmes)

Utrechtse Universiteitsbibliotheek (UBU) [14] Digital books and articles

Since many of the data provided by these organisations are highly relevant to humanities researchers, we want these data to be available via the CLARIN infrastructure. And they are, for NIBG, for UBU, or will soon be (KB).

4.1 Virtual Language Observatory

The Virtual Language Observatory (VLO) offers facilities for browsing and searching in CMDI resource descriptions. Once the desired resource descriptions have been found, links to the actual resources (data and software) enable the researcher to make use of the resources in his/her research.

The VLO enables a user to do a keyword (string) search for keywords that occur in the resource descriptions.When you type in a keyword, the VLO provides suggestions for keywords that occur in the resource descriptions (query completion).[15] In addition to keyword search, the VLO offers faceted browsing: one can select values for a range of facets such as *language, subject collection, format, resource type, organisation, continent, national project, country, keyword, modality, data provider* and *genre.* The VLO currently gives access to over 650k resource descriptions, and this number is expected to grow considerably in the coming

[11] This type of CLARIN Centre is currently only distinguished the Netherlands.

[12] National Library

[13] Netherlands Institute for Sound and Vision

[14] Utrecht University Library

[15] At the time of writing, only keywords from selected resource description fields were presented.

years.[16] Important resources that are relevant to the Netherlands are currently still absent, inter alia the resources for the Dutch language at the HLT-Agency (TST-Centrale), which aims become a certified CLARIN Centre for the CLARIN ERIC partner Dutch Language Union. For more information on finding data through the VLO, I refer to Van Uytvanck (2014).

4.2 Meertens CLARIN Metadata Search

The Meertens CLARIN Metadata Search Engine(Zhang *et al.* (2012) offers an alternative way to find resources through resource descriptions. This search engine operates in principle on the same resource descriptions as the VLO: the resource descriptions harvested for the VLO. But snapshots from the resource descriptions harvested for the VLO are taken a specific intervals, so there may be a difference between what is visible via the Meertens Metadata Search and the VLO.

The engine also offers keyword (string) search, and it offers query completion but now on all keywords that occur in the resource descriptions, and it also indicates in which resource description element the keyword occurs and how often. This helps in selecting the desired or most relevant resource descriptions. For example, after typing in the character sequence *pe*, suggested keywords starting with this character sequence are immediately shown, e.g. *period*, in combination with the information that it occurs 403 times in the *description* element of the resource description element *time coverage* (see left top corner of figure 5).

The interface also makes suggestions for other searches (see under *You could also look for...* in the mid right part of figure 5). Keywords suggested there form the most important keywords related to the query based on the TF-IDF statistics[17].

When a query has run, the search selection is automatically stored, so that a user can refine the search within the current collection. There is also an option to remove the whole search selection.

The interface offers different overviews of the retrieved results, inter alia a dynamic word cloud of the aggregated content within the resource description element (see mid left part of figure 5), and it offers different visualisations of the aggregated search features: resources for which a geo-reference is available are displayed on a map (see left bottom part of figure 5), and there are editable charts for displaying the date ranges of documents (see right bottom part of figure 5).

Finally, it recommends related resources (see figure 6) by providing links to related resource descriptions and a snippet of the first recommended resource description.

[16] However, this number does not say very much, because different providers of resource descriptions may have different views on the granularity of the resource descriptions: in some cases a resource description describes just one small piece of text (e.g. a newspaper article, or a song), in other cases it describes a full collection of newspaper articles for a whole year of a specific newspaper. Finding a good balance between the optimal granularity in function of the main purpose of the VLO (finding relevant research resources) will be a major challenge in the coming years.

[17] A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. See `http://en.wikipedia.org/wiki/Tf%E2%80%93idf`

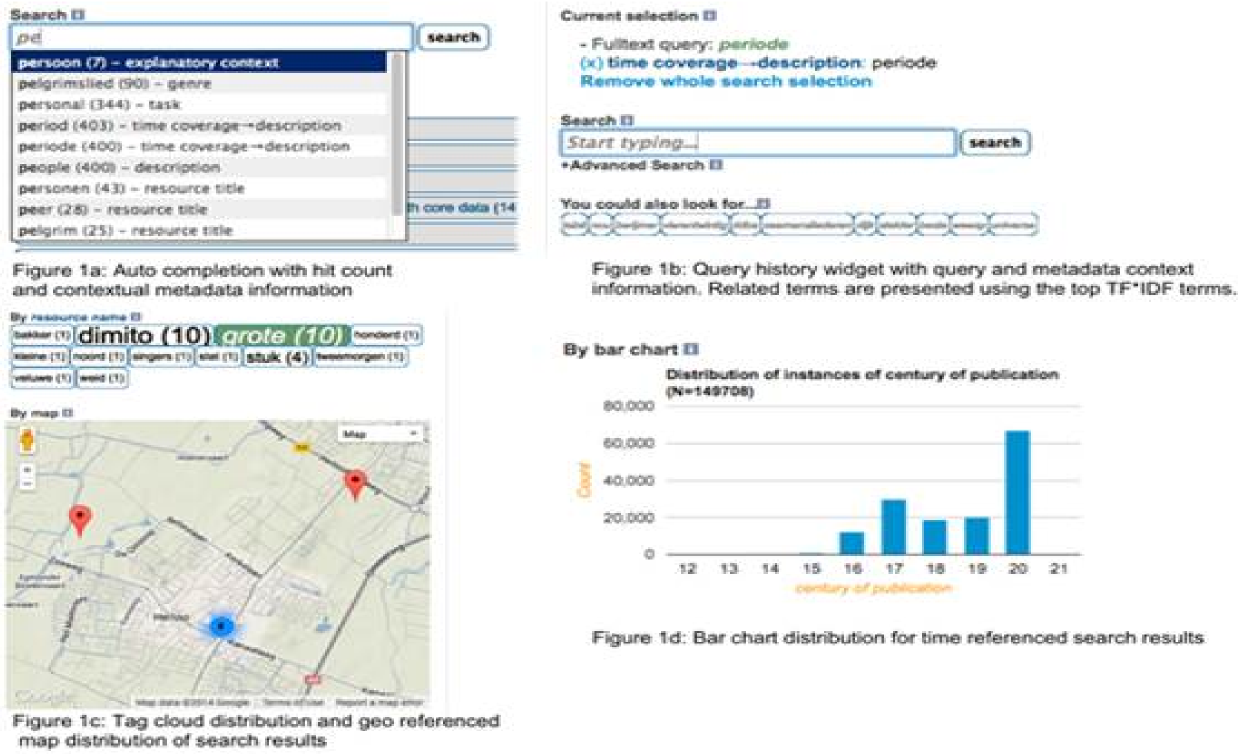

Search
pel
search
persoon (7) – explanatory context
pelgrimslied (90) – genre
personal (344) – task
period (403) – time coverage→description
periode (400) – time coverage→description
people (400) – description
personen (43) – resource title
peer (28) – resource title
pelgrim (25) – resource title
h core data (14
Current selection
- Fulltext query: periode
(x) time coverage→description: periode
Remove whole search selection
Search
Start typing...
search
+Advanced Search
You could also look for...
By resource name
bakker (1)
dimito (10)
grote (10)
honderd (1)
kleine (1)
noord (1)
singers (1)
stel (1)
stuk (4)
tweemorgen (1)
veluwe (1)
weid (1)
By map
Map
By bar chart
Distribution of instances of century of publication
(N=149708)
80,000
60,000
40,000
20,000
0
Count
12
13
14
15
16
17
18
19
20
21
century of publication

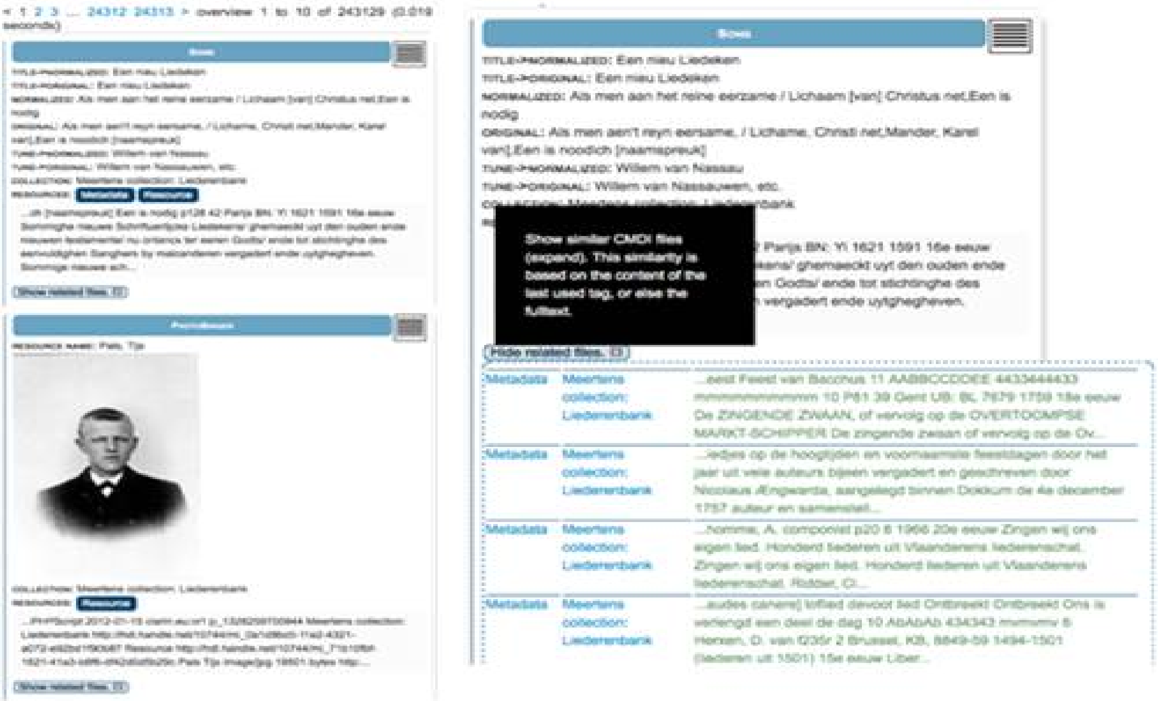
Figure 1a: Auto completion with hit count and contextual metadata information

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

Figure 5: Meertens CLARIN Metadata Search Interface



< 1 2 3 ... 24312 24313 > overview 1 to 10 of 243129 (0.019 seconds)

TITLE->NORMALIZED: Een nieu Liedeken
TITLE->ORIGINAL: Een nieu Liedeken
NORMALIZED: Als men aan het reine eerzame / Lichaam [van] Christus net,Een is nodig
ORIGINAL: Als men aen't reyn eersame, / Lichame, Christi net,Mander, Karel van],Een is noodich [naamspreuk]
TUNE->NORMALIZED: Willem van Nassau
TUNE->ORIGINAL: Willem van Nassauwen, etc.
COLLECTION: Meertens collection: Liederenbank
RESOURCES: Metadata Resource

...ch [naamspreuk] Een is nodig p128 42 Parijs BN: Yi 1621 1591 16e eeuw Sommighe nieuwe Schriftuerlijcke Liedekens/ ghemaeckt uyt den ouden ende nieuwen testamente/ nu ontrent ter eeren Godts/ ende tot stichtinghe des eenvoudighen Sanghers by malcanderen vergadert ende uytghegheven. Sommige nieuwe sch...

RESOURCE NAME: Pels, Tjo

COLLECTION: Meertens collection: Liederenbank
RESOURCES: Resource

...PHPScript 2012-01-15 clarin.eu:cr1:p_1328259700944 Meertens collection: Liederenbank http://hdl.handle.net/10744/mi_[illegible] [illegible] Resource http://hdl.handle.net/10744/mi_[illegible] Pels Tjo image/jpg 19501 bytes http:...

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

TITLE->NORMALIZED: Een nieu Liedeken
TITLE->ORIGINAL: Een nieu Liedeken
NORMALIZED: Als men aan het reine eerzame / Lichaam [van] Christus net,Een is nodig
ORIGINAL: Als men aen't reyn eersame, / Lichame, Christi net,Mander, Karel van],Een is noodich [naamspreuk]
TUNE->NORMALIZED: Willem van Nassau
TUNE->ORIGINAL: Willem van Nassauwen, etc.
COLLECTION: Meertens collection: Liederenbank

Show similar CMDI files (expand). This similarity is based on the content of the last used tag, or else the fulltext.

...2 Parijs BN: Yi 1621 1591 16e eeuw ...kens/ ghemaeckt uyt den ouden ende ...en Godts/ ende tot stichtinghe des ...n vergadert ende uytghegheven.

Metadata	Meertens collection: Liederenbank	...eest Feest van Bacchus 11 AABBCCDDEE 4433444433 mmmmmmmmmm 10 P81 39 Gent UB: BL 7679 1759 18e eeuw De ZINGENDE ZWAAN, of vervolg op de OVERTOOMPSE MARKT-SCHIPPER De zingende zwaan of vervolg op de Ov...
Metadata	Meertens collection: Liederenbank	...edjes op de hoogtijden en voornaamste feestdagen door het jaar uit vele auteurs bijeen vergadert en geschreven door Nicolaus Ængwarda, aangelegd binnen Dokkum de 4e december 1757 auteur en samensteil...
Metadata	Meertens collection: Liederenbank	...homme, A. componist p20 8 1966 20e eeuw Zingen wij ons eigen lied. Honderd liederen uit Vlaanderens liederenschat. Zingen wij ons eigen lied. Honderd liederen uit Vlaanderens liederenschat. Ridder, Cl...
Metadata	Meertens collection: Liederenbank	...audes canere] loflied devoot lied Ontbreekt Ontbreekt Ons is verlengd een deel de dag 10 AbAbAb 434343 mmvmmv 8 Horsen, D. van f235r 2 Brussel, KB, 8849-59 1494-1501 (liederen uit 1501) 15e eeuw Liber...

Figure 2b: Recommendation list of related results

Figure 6: Meertens CLARIN Metadata Search Interface: recommended resources

5 Applying software to data

There is a lot of software in the CLARIN infrastructure that can be applied to data. Even if we restrict attention to the Netherlands, there are too many to describe them all here. Instead, we will briefly describe what *types* of tools and services CLARIN currently contains.

The tools and services can be found most easily via the Netherlands CLARIN portal, under Services, or via the overview of data, tools, demonstrators and applications on the CLARIN-NL web site.

Three major classes of applications and services will be discussed: Searching in data (section 5.1, annotation and related tools (section 5.2), and processing data (section 5.3).

5.1 Searching in data

Federated Content Search is a technique in which a single query can be launched to search in multiple resources that are stored in a different locations and that each may have their own particular format. A limited form of federated content search is possible in data from CLARIN-NL, CLARIN-D and LINDAT (CLARIN-CZ) via the CLARIN-D Federated Content Search graphical user interface. This federated content search is limited in two respects: first, it currently only enables string (keyword) search, and second, it only applies to a limited number of resources in the CLARIN infrastructure.

There are also many search engines that apply to specific resources only. They include search engines for searching in lexical data (lexicons, lexico-semantic databases, databases of multi-word expressions, sign language lexicons, etc.), for searching in linguistically annotated data (text corpora, treebanks, audio-visual corpora, and typological databases), for searching in literary data and metadata (e.g. in literary works and descriptions of literary works), for searching in historical and contemporary textual and audio-visual corpora (e.g. ships records, historical and contemporary newspapers, tv and radio programmes, oral history collections, letters of 17th century scholars, parliamentary proceedings etc.), searching for religious data (e.g. pilgrimage reports, the Hebrew Bible), as well as searching for migration patterns within the Netherlands.

Most of these search engines also have facilities for analyzing the found data and to visualize them in various ways.

5.2 Annotation and Related Tools

A number of tools focus on annotating resources, i.e. enriching them with new information. They include tools for diarisation of audio files[18], for annotating time-based resources (audio, video) and multimodal resources, as well as for evaluating the quality of phonetic transcriptions of speech files.

Several of the tools described in section 5.3 can also be used for annotation purposes.

[18] i.e. annotating where in an audio file there is speech (instead of other sounds), and identifying who is speaking in the parts containing speech

5.3 Processing data

Tools for processing data include a tool for orthographic normalisation, which is also embedded in a work flow for converting digital images into textual resources in TEI format[19], a tool chain and methodology for converting legacy data sets in the area of maritime history, an application to analyze writing style, and an application for the analysis and visualisation of dialect variations.

It also includes tools for tokenizing, lemmatizing, part-of-speech tagging and parsing for mediaeval and for modern Dutch, tools for assigning semantic roles, co-reference relations, and for identifying and analyzing named entities, and tools for the automatic orthographic transcriptions of the speech in audio files. Most of these have been implemented as web services or as work flows of web services.

6 Storing data and software in CLARIN

If you have or are going to create a resource, e.g. in the context of a research project, you can store this resource in the CLARIN infrastructure, and you are strongly recommended to do so.

I will first discuss why it makes sense to store your resource in CLARIN (section 6.1). Next, I will describe how you should store a resource in CLARIN, initially focusing on new resources. When you want to store a resource in CLARIN, two parties are involved: you, as the data provider, and a CLARIN centre. I will describe what you have to do (section 6.3) and what the CLARIN centre has to do (section 6.4), initially for new resources. Finally, I will discuss what has to be done for resources that already exist (section 6.5).

6.1 Why should I store my resource in CLARIN?

The first question that arises when you have resource: why would I store it in the CLARIN infrastructure?

Well, there are many reasons.[20] A very important one is that you may benefit from doing so: if you make your resource ready for storage in CLARIN, you may easily make use of existing software and data in CLARIN, so that you can produce your data or software more efficiently, with better quality and/or with more features. You may also use CLARIN tools such as search engines, analysis tools, and visualisation tools on your resources, so that you can use your resource immediately in your research. And when your resource is in the CLARIN infrastructure, you can be sure it is stored safely, always easily accessible, and you do not have to worry about these data in a world where software updates and upgrades are frequent so that your resources can become obsolete in a very short period of time. It often happens that researchers change research topics and do not need research data made in an earlier project in the next project. However, when you do need your resource in a later stage, you do not have to worry where it is, and whether the medium you stored it on is still working: you can be sure to find it and get access to it via CLARIN.

[19] TEI (Text Encoding Initiative) is a widely used standard for encoding textual resources supported by CLARIN.

[20] Here's a clip by DANS on the importance of data sharing(in Dutch).

A second reason is that others may benefit from your resource. There are always unexpected uses of research data, either immediately or in some cases only years or even decades later. CLARIN ensures that all researchers have access to the resources used in or resulting from research.

Furthermore, making your resource available via CLARIN fits in well with the general scientific attitude of openness. Most resources are produced with public money, so it is important that the whole society can benefit from the resource you produced..

There are also reasons of integrity: we have recently encountered several scandals in the Netherlands where faked data were used in research. Making resources openly available via CLARIN will reduce the risks of such fraud. More generally, science progresses by being open to criticism, and verification and replication of research results are important instruments to make progress in science and are essential for the proper conduct of science: visibility and accessibility of your research data and software is essential for that, and CLARIN provides ideal facilities for that.

Since openness about your research data and results is an essential ingredient for the proper conduct of science, more and more scientific journals are beginning to require that you publish your research data and software, so that your results are verifiable and replicable. For the same reasons, also funding agencies are beginning to require an explicit data management plan, so that data produced in a research project do not get lost after the research project has finished[21] and are available for verification and replication purposes.[22] Workshops to elaborate this policy are being organized.[23] The Standard Evaluation Protocol (SEP) 2015-2021 by VSNU, KNAW and NWO (VSNU *et al.* (2014)) states that the assessment committee "is interested in how the unit deals with research data, data management and integrity" (p. 9) and the self-evaluation should describe "how the unit deals with and stores raw and processed data" (p. 23). So, also your own research unit will most likely require that you deal carefully with data: CLARIN offers the facilities for that.

6.2 How to store resources in CLARIN

If your research will lead to new resources, it is important to start immediately taking into account that they will be stored in the CLARIN infrastructure. Ideally, you start with this before any data or software have been produced. If you have already produced part or all of your resource, see section 6.5.

Two parties are involved in storing resources in CLARIN: you, as resource provider, and a CLARIN Centre. Both parties have responsibilities when a resource has to be stored in CLARIN. We describe these responsibilities in separate sections: the responsibilities of the resource provider in section 6.3, the responsibilities of the CLARIN Centre in section 6.4.

It is important to contact a CLARIN Centre as soon as possible. The CLARIN centre will be able to help you with preparing your resource for incorporation in CLARIN, and your resource must be stored at a CLARIN centre for it to become part of the CLARIN infrastructure.

[21] which, unfortunately, has happened a lot in the past!

[22] See for example NWO (2014:19, article 30).

[23] For example, a symposium on data management and open access on September 10, 2014, organized by the Dutch funding agency NWO and SURFacademy.

CLARIN Centres come in different types.[24] The type relevant in this context is type B.

The Netherlands has multiple Type B CLARIN centres. They include Meertens Institute (Amsterdam), the Language Archive (TLA) of the Max Planck Institute for Psycholinguistics (MPI, Nijmegen), Huygens ING Institute (The Hague), Institute for Dutch Lexicology (INL, Leiden), and Data Archiving and Networked Services (DANS, The Hague). Which one to choose? Well, that depends on the type of resource you have and its primary intended research use. The CLARIN Portal provides information about the various centres and the type of resources they are most suited for. Similar information can be found on the CLARIN-NL website Centres Page.

Here's a brief overview of the Dutch Type B CLARIN Centres and the resource types they are most suited for:

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) resources related to the lexicological study of the Dutch language.

Data Archiving and Networked Centres (DANS) digital research data generally.

6.3 What you must do

The first thing to do is to define clearly what your resource is going to be. Once this is clear, you can select a CLARIN centre, and contact them.[25] Next, you have to ensure that legal and ethical issues do not prevent you from incorporating the resource in the CLARIN infrastructure and making them available to other researchers. There are several ways of doing this, depending on the type of resource. If the owner of the resource is not you, you will have to obtain explicit permission for this through some license agreement. If subjects participate in a resource creation project, you will have to ask them explicit permission to use the resource in the CLARIN infrastructure. The CLARIN centre can help you with this, and there are templates for license agreements and for consent forms. Together with the Centre you will have to ensure that ethical issues (mostly privacy issues), where they arise, are properly dealt with.

We will discuss what you have to do, initially focusing on data. We dedicate a separate paragraph to the case where your resource is software.

CLARIN-recommended formats You have to determine a CLARIN-recommended format for your resource. A list of CLARIN-recommended formats, protocols, etc can be found

[24] This document contains an overview of the different type of CLARIN Centres.

[25] Contact information can be found here.

here. Again, consult with the CLARIN Centre on this issue, or ask help from the CLARIN-NL helpdesk (`helpdesk@clarin.nl`). Since we are in the area of research, it is possible that your resource is of a completely new type, for which no CLARIN-recommended formats exist. It is also possible that none of the CLARIN-recommended formats can accommodate all elements of your resource, even though your resource is not of a completely novel type. In all these cases, consult with the CLARIN-NL helpdesk first before continuing.

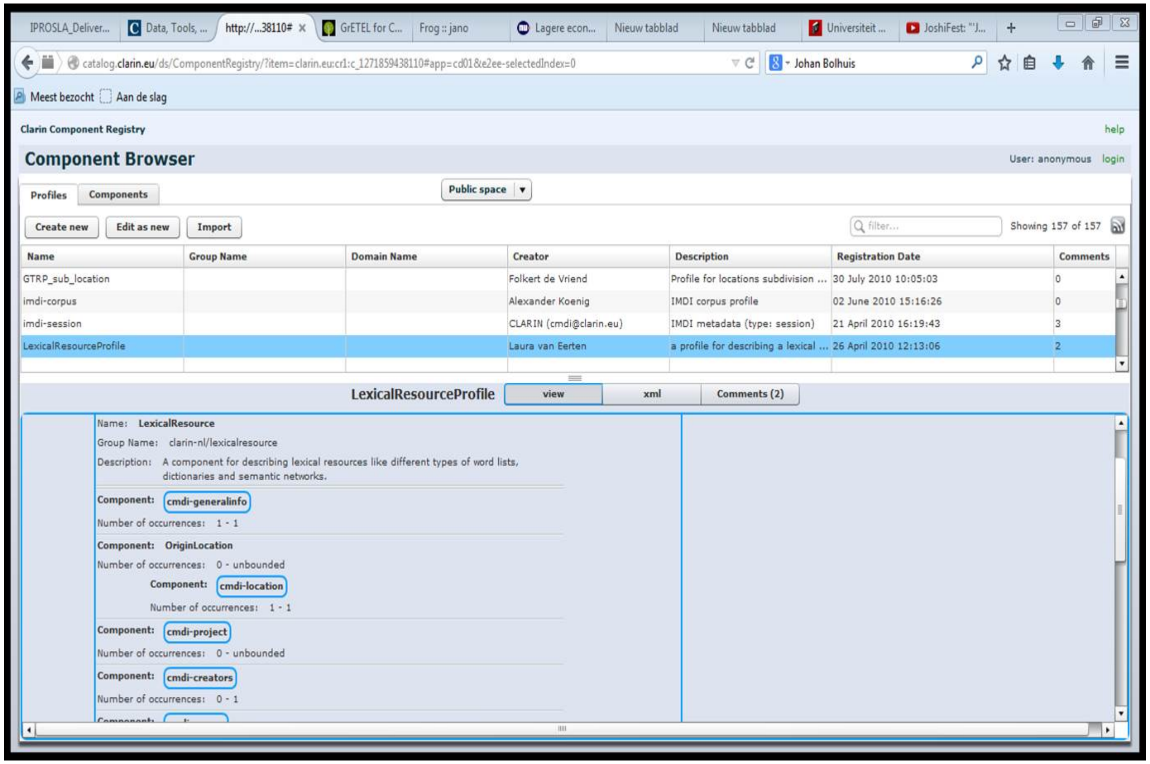
Resource Descriptions One or more descriptions must be made of your resource. These resource descriptions (also called 'metadata') must be in CMDI-format (Broeder *et al.* (2010)). CMDI (Component-based Metadata Infrastructure) provides a model for resource descriptions, and a format for resource descriptions. It also provide tools to make resource descriptions . CMDI resource descriptions are written in XML (eXtensible Markup Language). CMDI does NOT in any way proscribe the contents of the resource descriptions. That is completely up to you (though CMDI helps you in several ways to create correct and 'useful' resource descriptions.

CMDI resource descriptions are structured in accordance with a *profile*. A profile describes which elements can or must be used in a resource description. Resource description elements are XML elements, consisting of a *name*, a *value*, and a (possibly empty) set of attribute-value pairs. An example CMDI element is illustrated in figure 9. Often, a group of such elements naturally belong together, e.g because together they describe a particular aspect of a resource. One can group such elements in a resource description *component*. This enables you to treat such a collection of resource description elements as a unit. Resource description components consist of a combination of components and resource description elements. An example CMDI component is illustrated in figure 8. A profile consists of a combination of components and elements. An example profile is illustrated in figure 7.

This component-based system provides high flexibility: *you* can determine the contents of the descriptions for your resource by defining your own profiles, components, and elements. CMDI helps you with this in a variety of ways:

- A list of existing profiles and components enables you to reuse what has already been made by others: it this saves you work, and you can profit from work done by others.
- A profile and component editor [login required] enables you to create your own profiles and components if existing profiles and components are not suited to you.
- A metadata editor: ARBIL enables you to create descriptions for your resources in accordance with the selected profile in an easy and user-friendly manner.

The flexibility offered by CMDI also has some drawbacks. One has to be aware that one major purpose of these resource descriptions is the discovery of the resources by others. It is therefore important to include information that characterizes this resource and distinguishes it from other resources. It is therefore also highly recommended to use certain components that contain important resource description elements you are likely to overlook if you have to make your profile from scratch (e.g. the GeneralInfo component, which contains elements for general information about the resource, e.g. its name, title, the time coverage of the data, etc.). It is also important to be aware that certain properties that are 'obvious' to you are not obvious to other researchers and must therefore be included in a proper resource



catalog.clarin.eu/ds/ComponentRegistry/?item=clarin.eu:cr1:c_1271859438110#app=cd01&e2ee-selectedIndex=0

Clarin Component Registry

Component Browser

User: anonymous

Showing 157 of 157

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
GTRP_sub_location			Folkert de Vriend	Profile for locations subdivision ...	30 July 2010 10:05:03	0
imdi-corpus			Alexander Koenig	IMDI corpus profile	02 June 2010 15:16:26	0
imdi-session			CLARIN (cmdi@clarin.eu)	IMDI metadata (type: session)	21 April 2010 16:19:43	3
LexicalResourceProfile			Laura van Eerten	a profile for describing a lexical ...	26 April 2010 12:13:06	2

LexicalResourceProfile

Name: LexicalResource
Group Name: clarin-nl/lexicalresource
Description: A component for describing lexical resources like different types of word lists, dictionaries and semantic networks.

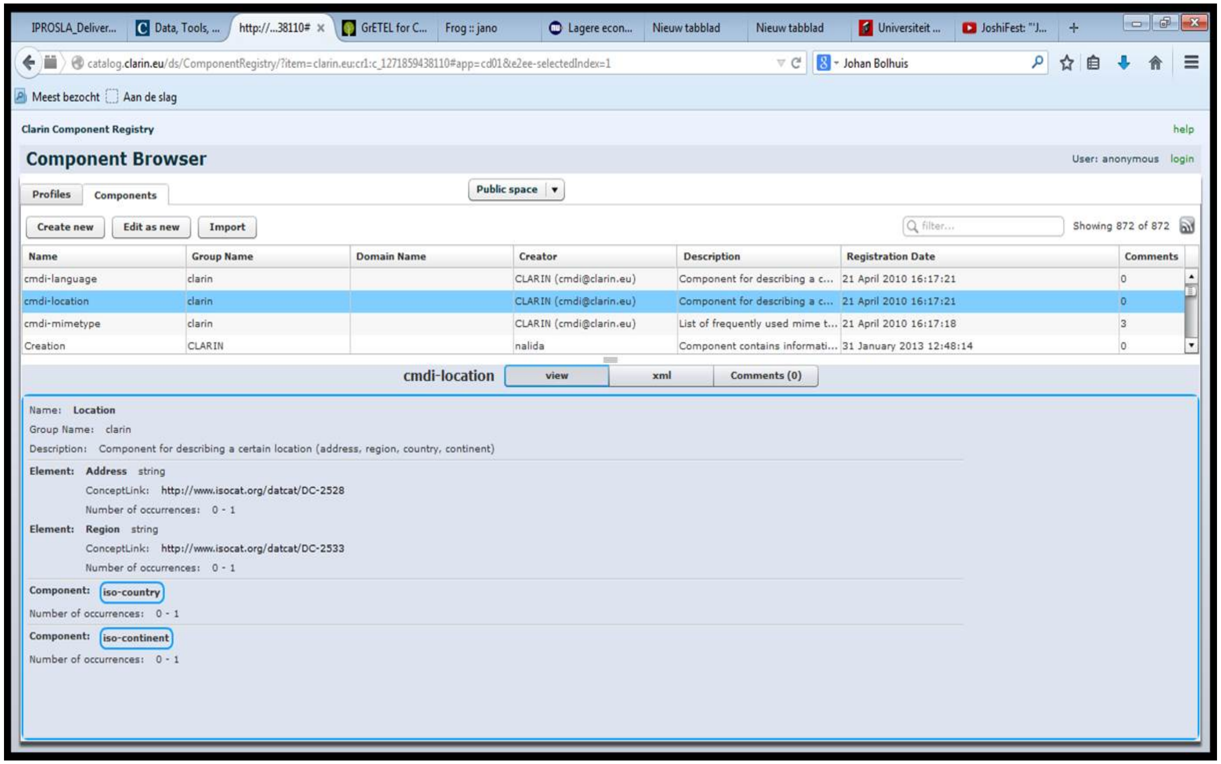
Component: cmdi-generalinfo
Number of occurrences: 1 - 1

Component: OriginLocation
Number of occurrences: 0 - unbounded
Component: cmdi-location
Number of occurrences: 1 - 1

Component: cmdi-project
Number of occurrences: 0 - unbounded

Component: cmdi-creators
Number of occurrences: 0 - 1

Figure 7: CMDI Profile example



Clarin Component Registry

Component Browser

User: anonymous

Profiles | Components | Public space

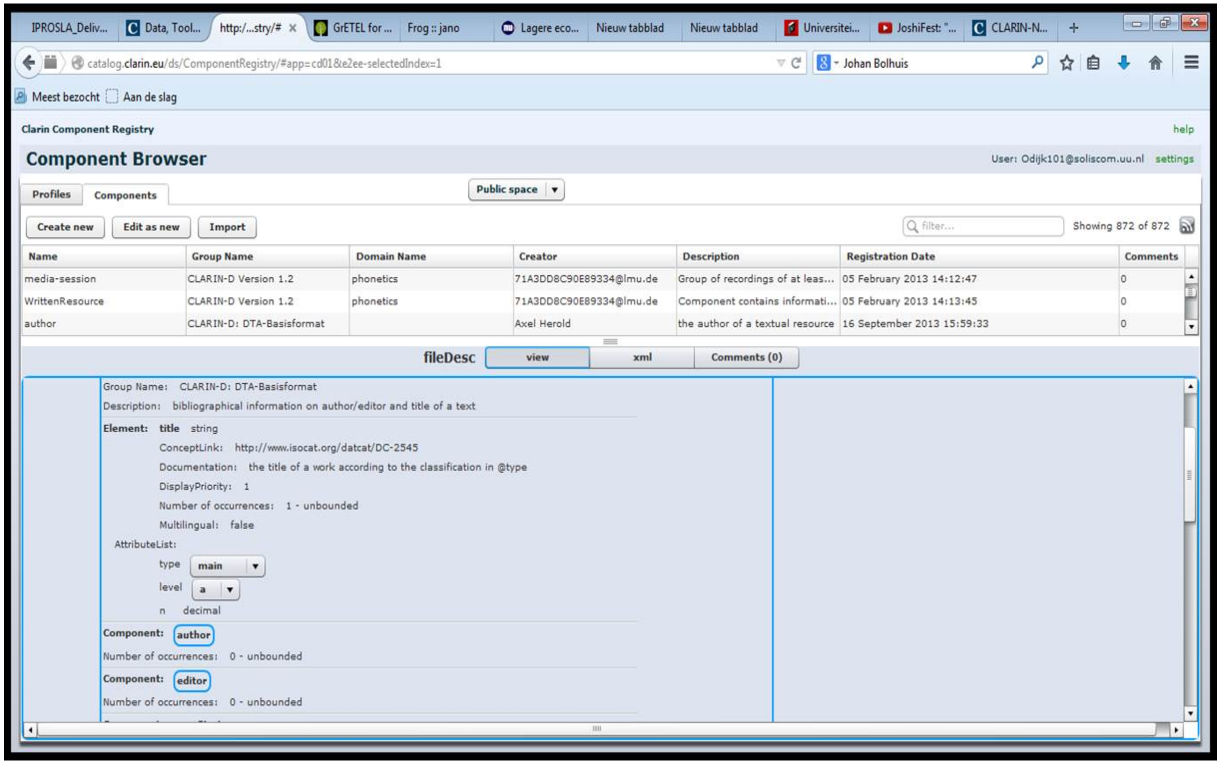
Create new | Edit as new | Import | filter... | Showing 872 of 872

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
cmdi-language	clarin		CLARIN (cmdi@clarin.eu)	Component for describing a c...	21 April 2010 16:17:21	0
cmdi-location	clarin		CLARIN (cmdi@clarin.eu)	Component for describing a c...	21 April 2010 16:17:21	0
cmdi-mimetype	clarin		CLARIN (cmdi@clarin.eu)	List of frequently used mime t...	21 April 2010 16:17:18	3
Creation	CLARIN		nalida	Component contains informati...	31 January 2013 12:48:14	0

cmdi-location | view | xml | Comments (0)

Name: Location
Group Name: clarin
Description: Component for describing a certain location (address, region, country, continent)
Element: Address string
ConceptLink: http://www.isocat.org/datcat/DC-2528
Number of occurrences: 0 - 1
Element: Region string
ConceptLink: http://www.isocat.org/datcat/DC-2533
Number of occurrences: 0 - 1
Component: iso-country
Number of occurrences: 0 - 1
Component: iso-continent
Number of occurrences: 0 - 1

Figure 8: CMDI Component example



Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
media-session	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Group of recordings of at leas...	05 February 2013 14:12:47	0
WrittenResource	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Component contains informati...	05 February 2013 14:13:45	0
author	CLARIN-D: DTA-Basisformat		Axel Herold	the author of a textual resource	16 September 2013 15:59:33	0

Group Name: CLARIN-D: DTA-Basisformat
Description: bibliographical information on author/editor and title of a text
Element: **title** string
ConceptLink: http://www.isocat.org/datcat/DC-2545
Documentation: the title of a work according to the classification in @type
DisplayPriority: 1
Number of occurrences: 1 - unbounded
Multilingual: false
AttributeList:
type main ▾
level a ▾
n decimal
Component: **author**
Number of occurrences: 0 - unbounded
Component: **editor**
Number of occurrences: 0 - unbounded

Figure 9: CMDI Element example

description. For example, several researchers that only work with the Dutch language have omitted an indication of the language of the resource in their resource description. The same holds for the *resource type* element, which was omitted by researchers who mainly work with text corpora. It is also important to give your resource a name: that makes referring to it much easier. And use explicit versioning from the start: otherwise it will become very difficult to know later which version is intended.

Reusing existing profiles and components will help you get better resource descriptions, and you do not have to reinvent wheels that already exist. It is strongly advised to follow an introductory course on CMDI. They are held regularly in the Netherlands. For any questions on CMDI, you can also contact `cmdi@clarin.eu`.

Explicit semantics The flexibility of CMDI has other consequences as well. In rigid resource description schemes (e.g a CSV format), the position of an element determines its interpretation and in certain schemes (e.g http://dublincore.org/) the names of elements and their values are proscribed. But with CMDI, you can choose your own profiles, components and resource description elements, give resource description elements any name you like, and you can also choose the labels for the values of these elements. But then how does another rersearcher, or a computer programme 'know' what you mean with it?

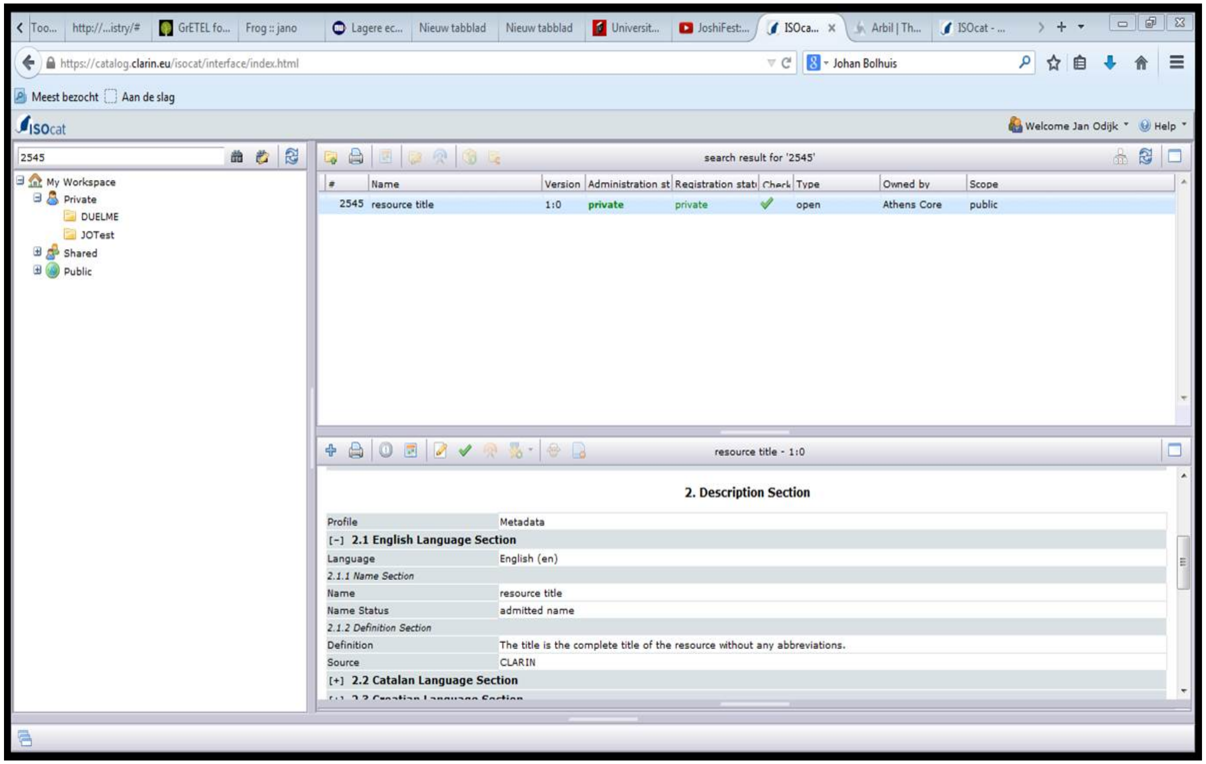
The flexibility offered by CMDI requires explicit semantics! The CLARIN infrastructure must 'know what you mean with your resource description elements, otherwise it cannot use faceted browsing in the VLO or the Meertens Metadata Search Engine.

Explicit semantics for a resource or resource description is obtained by explicitly linking each element and its possible values in the resource and resource description to an element of a CLARIN-recognized concept or data category registry. The most prominent data category registry in CLARIN is ISOCAT (Kemps-Snijders *et al.* (2010)). Figure 10 illustrates a specific data category in ISOCAT. An example of an explicit link to an ISOCAT data category in a CMDI element definition can be seen in figure 9 after *Concept Link*.

There is much more to be said about ISOCAT as a registry for data categories. For example, ISOCAT is basically just a flat list of data categories. However, often it is desired to specify the relation between data categories. This can be done in a special registry, called RELCAT (which currently only exist in an α-version). Furthermore, it is sometimes necessary or convenient to know more about the internal structure of a resource. For that purpose, the registry SCHEMACAT (α-version) has been set up. Finally, ISOCAT may be the primary concept registry in CLARIN, it is not the only one. For certain types of information ISOCAT is not particularly suited (e.g. names of organisations in all their variants), for others independent registries exist and are maintained (e.g. for language codes: ISO639-3, maintained by ISO). In order to use such other registries in addition to ISOCAT in a transparent manner, the CLAVAS Vocabulary Service has been set up as an interface to data category registries and vocabularies. CLAVAS currently provides access to three vocabularies:

1. ISO-639-3 language codes, as published by the Summer Institute of Linguistics (SIL).
2. Closed and simple Data Categories from the ISOcat metadata profile
3. A manually constructed and curated list of Organization names[26], based on the CLARIN

[26] requires login



#	Name	Version	Administration st	Registration stat	Check	Type	Owned by	Scope
2545	resource title	1:0	private	private	✓	open	Athens Core	public

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	
[+] 2.3 Croatian Language Section	

Figure 10: Example Data Category in ISOCAT.

VLO. The curation of the names has been carried out by the CLARIN-NL Data Curation Service.

This paper can give no more than some initial pointers of what is available. In order to really use the registries and tools offered effectively, one has to attend dedicated tutorials on CMDI and ISOCAT. These have been and will be regularly organized in the Netherlands (each has been organized about twice a year over the past 5 years). Usually, the CLARIN Centre helps you in creating the CMDI metadata and the explicit semantics that it requires.

Live version v. exchange/archive version In several cases, data come in two versions: a version intended for exchange and for long term preservation (exchange/archive version), and a version that is actually used in services ('live version'). A concrete example is a lexicon: a CLARIN-supported format for lexicons is the Lexical Markup Framework (LMF). LMF-compatible text formats often make use of XML, and are excellently suited for exchange of data and for long term preservation (storage in an archive). However, this format is less suited for actual use by a service. For example, a simple search programme will operate unacceptably slowly if it has to work directly with the LMF textual format. Typically, the data have to be transformed into different formats, enriched with indexes, etc. for such a search service to operate in an acceptable way. This creates the problem that it must be ensured that the 'live' version and the exchange version must be kept consistent. This requires explicit versioning, and ideally the 'live' version is derived from the exchange version in a fully automated manner. If this is or might be the case for your data, consult with the CLARIN centre on how to deal with it.

Software Your resource may be software. Software comes in many varieties. First, software may run locally on a single desktop, or over the web. Second, software may have a user interface for specialists (e.g. a command-line interface), or an interface specifically designed for a specific user community (an *application*), or it may have an interface to other software (a (software) *service*).

Software intended for the CLARIN user group must of course have a dedicated interface. It preferably works over the web so that no software needs to be downloaded and installed. Such software thus typically comes in the form of a *web application*. For certain cases (e.g. language documentation field work), there is no or very limited internet availability, and a web application is not so useful: for such cases *desktop applications* are more suited.[27]

It is good practice to separate the programme that implements the interface from the backend software that provides the core functionality of the application. This backend may contain a single software programme but it might also contain multiple programmes that work together to provide the application's functionality. These programmes communicate with one another and therefore they are (software) *services*. For services that work over the web there are special protocols to make this communication possible. The ones supported in CLARIN are SOAP and REST. If you have a desktop programme, you will often want to turn

[27] There may be other considerations to prefer desktop over web applications, e.g. web interfaces are generally quite primitive and generally slow; if a sophisticated and/or fast operating interface is required, a desktop might be preferable. Ideally of course, one single interface operates both over the web and locally, and uses synchronisation/replication mechanisms to keep the local version and the version on the server in sync.

it into a web service in the CLARIN context. For this purpose, a special piece of software has been developed, called Computational Linguistics Application Mediator (CLAM), which turns your desktop software into a web service using the REST protocol. Though CLAM creates a web service, it actually also creates a simple web interface (hence a web application), but that is not necessarily the best interface for the targeted user group.

A piece of software is a resource, and therefore there must be a resource description for each piece of software.[28] A CMDI profile for the description of software exists and is further refined. Consult with the CLARIN centre on this.

This concludes the section on the tasks of the resource provider. We now turn to the tasks of the CLARIN Centre.

6.4 What the Centre must do

The first task of the CLARIN centre is to assist you with your tasks. They have experience with CMDI, with semantic interoperability, with IPR and ethical issues, with CLARIN-supported formats and protocols, so they are excellently suited for advising you in such matters.

Storing the resources The second task of the CLARIN Centre is to store your resource in its repository. Some centres use special software for this, e.g. LAMUS is used by MPI/The Language Archive, and EASY by DANS. The centre must of course also make the resource available and accessible in the CLARIN infrastructure for other researchers.

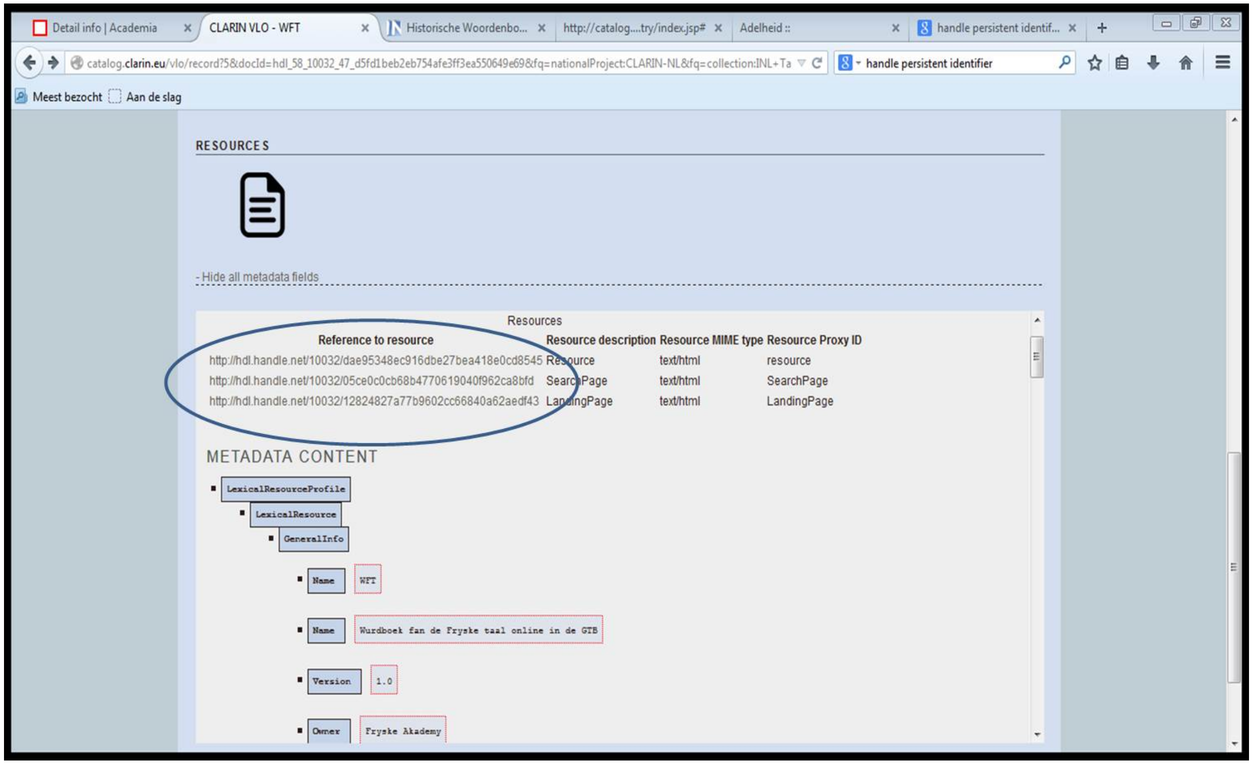
Resource description harvesting This is done through the resource description of the resource. The centre makes the resource descriptions of the resource available for harvesting by others through OAI-PMH.[29] Links to the actual resource are included in the resource description in the form of persistent identifiers, and the resource descriptions themselves are also assigned a persistent identifier (PID, see section 4).

Persistent Identifiers Each centre runs or uses a service for the assignment and resolution of persistent identifiers, i.e a system that assigns a PID when requested and associates it to a precise location, and that, given a PID, determines the precise location of the associated resource or resource description. In CLARIN, the so-called Handle system is used for the assignment and resolution of persistent identifiers. Figure 11 shows some examples of *Handle* PIDs in a CMDI resource description. These PIDs are preceded by the prefix `http://hdl.handle.net/`, which turns them into *actionable* PIDs, i.e, PIDs that are resolved and lead you to the resource it links to by clicking on it, just as any URL.

Legal and ethical restrictions The centre must also make provisions for legal and ethical restrictions, so that only persons who are allowed to actually get access to resources that have such restrictions. CLARIN aims to make available the resources as openly and with as little restrictions as possible. However, there are resources with legal and/or ethical restrictions, and therefore it is sometimes not possible to access such resources directly. The

[28] The term ‘metadata’ sounds somewhat odd for descriptions of software.

[29] Open Archives Initiative Protocol for Metadata Harvesting.



Detail info | Academia
CLARIN VLO - WFT
Historische Woordenbo...
http://catalog....try/index.jsp#
Adelheid ::
handle persistent identif...
catalog.clarin.eu/vlo/record?5&docId=hdl_58_10032_47_d5fd1beb2eb754afe3ff3ea550649e69&fq=nationalProject:CLARIN-NL&fq=collection:INL+Ta
handle persistent identifier
Meest bezocht
Aan de slag
RESOURCES
- Hide all metadata fields

Resources

Reference to resource	Resource description	Resource MIME type	Resource Proxy ID
http://hdl.handle.net/10032/dae95348ec916dbe27bea418e0cd8545	Resource	text/html	resource
http://hdl.handle.net/10032/05ce0c0cb68b4770619040f962ca8bfd	SearchPage	text/html	SearchPage
http://hdl.handle.net/10032/12824827a77b9602cc66840a62aedf43	LandingPage	text/html	LandingPage

METADATA CONTENT

- LexicalResourceProfile
  - LexicalResource
    - GeneralInfo
      - Name WFT
      - Name Wurdboek fan de Fryske taal online in de GTB
      - Version 1.0
      - Owner Fryske Akademy

Figure 11: Examples of PIDs in a CMDI resource description

restrictions can lead to various consequences: (1) a login may be required; (2) approving special usage conditions may be required; (3) signing a special license agreement may be required.

Logging in Hiding resource behind a login is intended, in the CLARIN context, to ensure that the user is an academic researcher, or has otherwise received special permission to access the relevant resources. There are also other reasons why login is sometimes necessary or desirable. For example, certain centres preserve data for you that you have uploaded to apply a service to it, as well as the data that result from this service. In such a case you want to make sure that only you (or your research team members) see and can manipulate these data, and you do not want to be bothered by data that belong to other researchers or research groups. Logging in is an essential ingredient of the means for achieving this. Certain services require a lot of computational resources, and the CLARIN centre where such a service runs wants to monitor its usage and to control the computational resources made available to you. Again, this requires logging in.

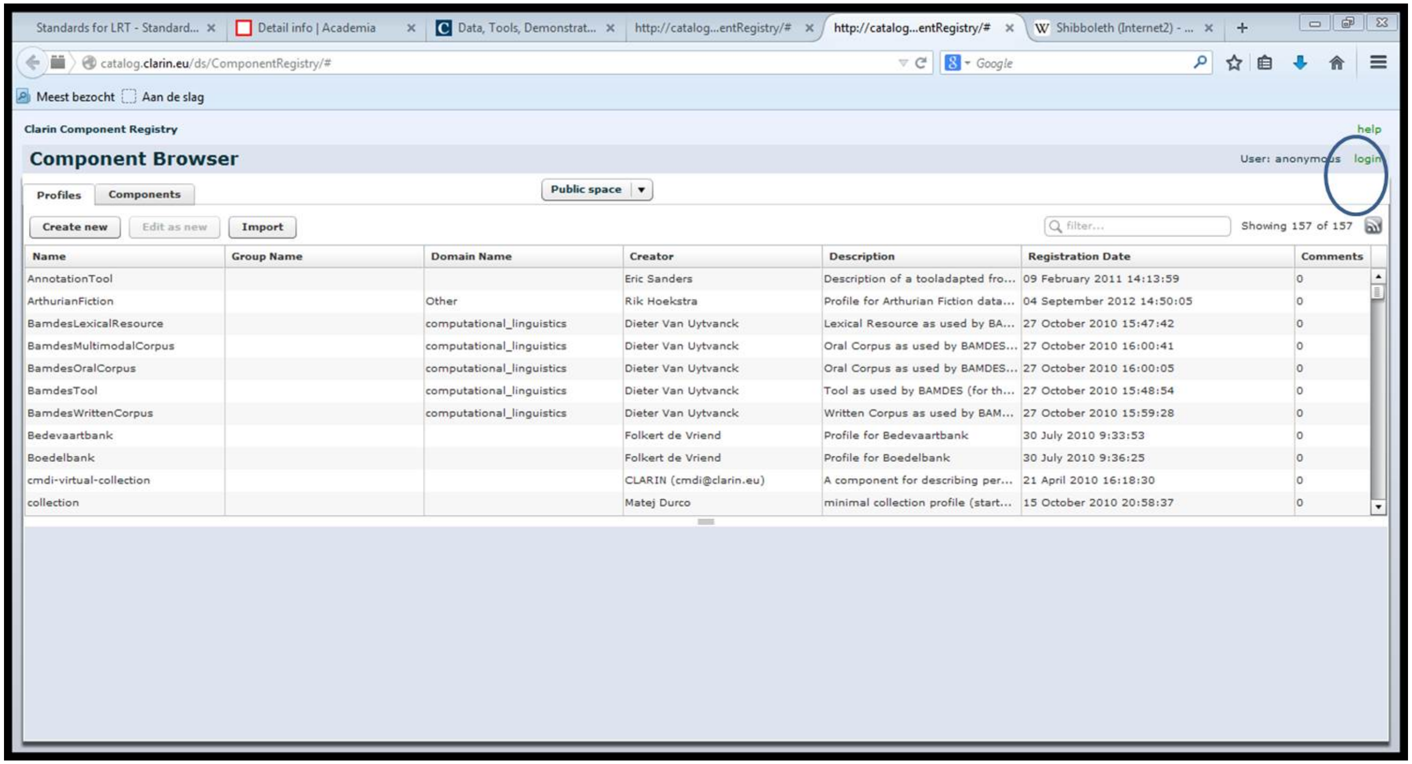
Logging in in the CLARIN infrastructure is not an obvious thing. The CLARIN infrastructure is a distributed infrastructure, so how can it be avoided that you have to login again each time a resource happens to be located at a different centre? How can it be avoided that you have to remember many different user names and passwords? And from the CLARIN centres perspective, how can it be avoided that each CLARIN Centre has to securely store user names, passwords and possibly other privacy-sensitive information?

Systems that take care of login and related matters are called *Authentication and Authorization Infrastructures (AAI)*: they *authenticate* you (determine who you are) and *authorize* you to do some things but not others. The AAI-system used in CLARIN is Shibboleth, and it avoids the problems mentioned above.

It works as follows:

- When you log in (for example, to edit a CMDI component in the CLARIN Component Registry, which requires login, see figure 12), you are directed to a login with your own institute. See figure 13.
- You then log in with you institute's user name and password. See figure 14.
- If the login is successful, the institute server 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 password. See figure 15.
- If you now go to another part of the CLARIN infrastructure that requires login (e.g the Adelheid web application), it 'knows' that you are already logged in, so you do not have to do this again: therefore this is called *Single Sign On (SSO)*. See figure 16.

Logging out is not so well-defined in this Single Sign On system. If you have logged in to a CLARIN service, and then go to second one (where no login is needed because the system 'knows' that you are logged in), you can try to log out of the first service, but then you are still logged in to the second service. So if you now go the first service again, you do not have to login despite having logged out, because it is a 'Single Sign On' system. Logout can only be achieved by closing all CLARIN services, and closing the browser(s) you used to access the CLARIN services.


Standards for LRT - Standard...
Detail info | Academia
Data, Tools, Demonstrat...
http://catalog...entRegistry/#
http://catalog...entRegistry/#
Shibboleth (Internet2) - ...
catalog.clarin.eu/ds/ComponentRegistry/#
Google
Meest bezocht
Aan de slag
help
login
Profiles
Components
Public space
Create new
Edit as new
Import
filter...

Clarin Component Registry

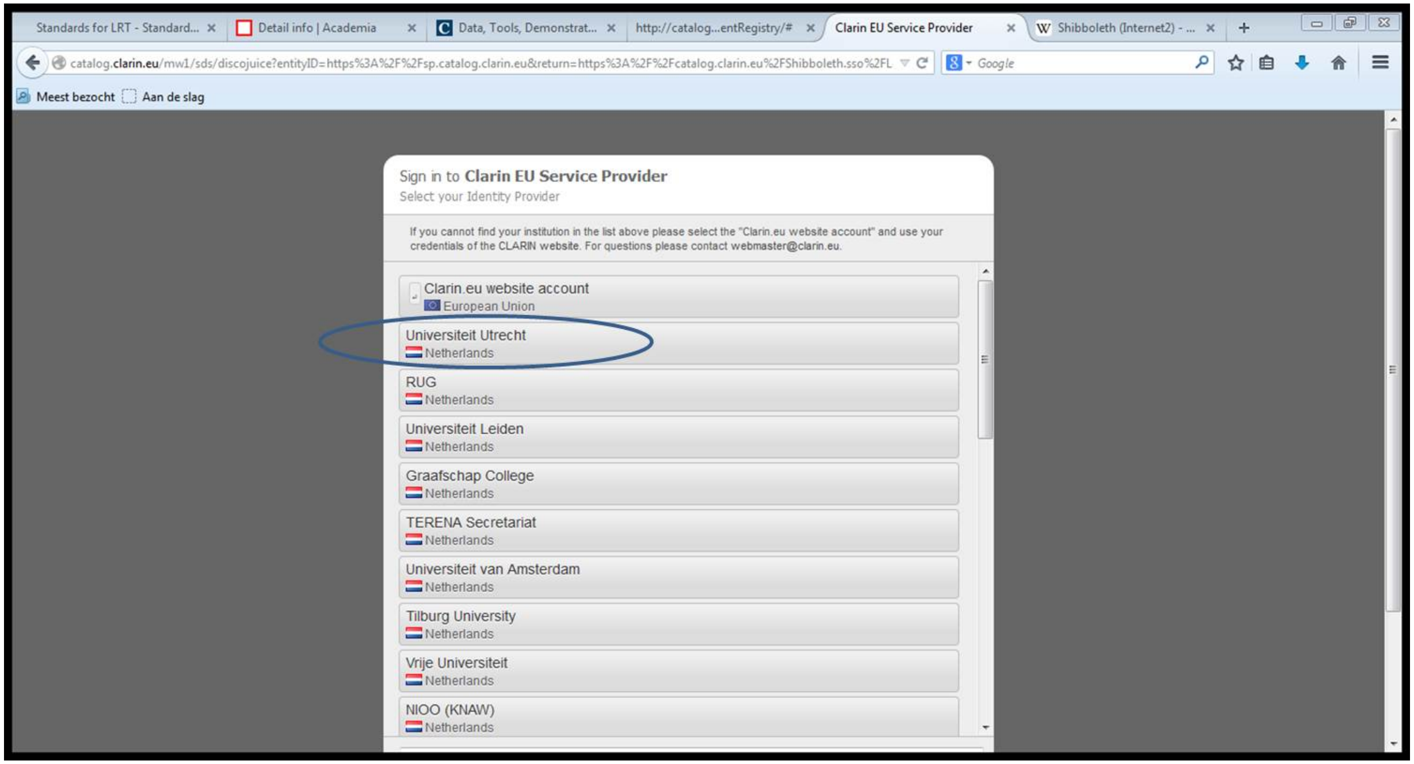
Component Browser

User: anonymous

Showing 157 of 157

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

Figure 12: You want to login in the CLARIN Component Registry.


Standards for LRT - Standard...
Detail info | Academia
Data, Tools, Demonstrat...
http://catalog...entRegistry/#
Clarin EU Service Provider
Shibboleth (Internet2) - ...
catalog.clarin.eu/mw1/sds/discojuice?entityID=https%3A%2F%2Fsp.catalog.clarin.eu&return=https%3A%2F%2Fcatalog.clarin.eu%2FShibboleth.sso%2FL
Google
Meest bezocht
Aan de slag

Sign in to **Clarin EU Service Provider**
Select your Identity Provider

If you cannot find your institution in the list above please select the "Clarin.eu website account" and use your credentials of the CLARIN website. For questions please contact webmaster@clarin.eu.

Clarin.eu website account
European Union

Universiteit Utrecht
Netherlands

RUG
Netherlands

Universiteit Leiden
Netherlands

Graafschap College
Netherlands

TERENA Secretariat
Netherlands

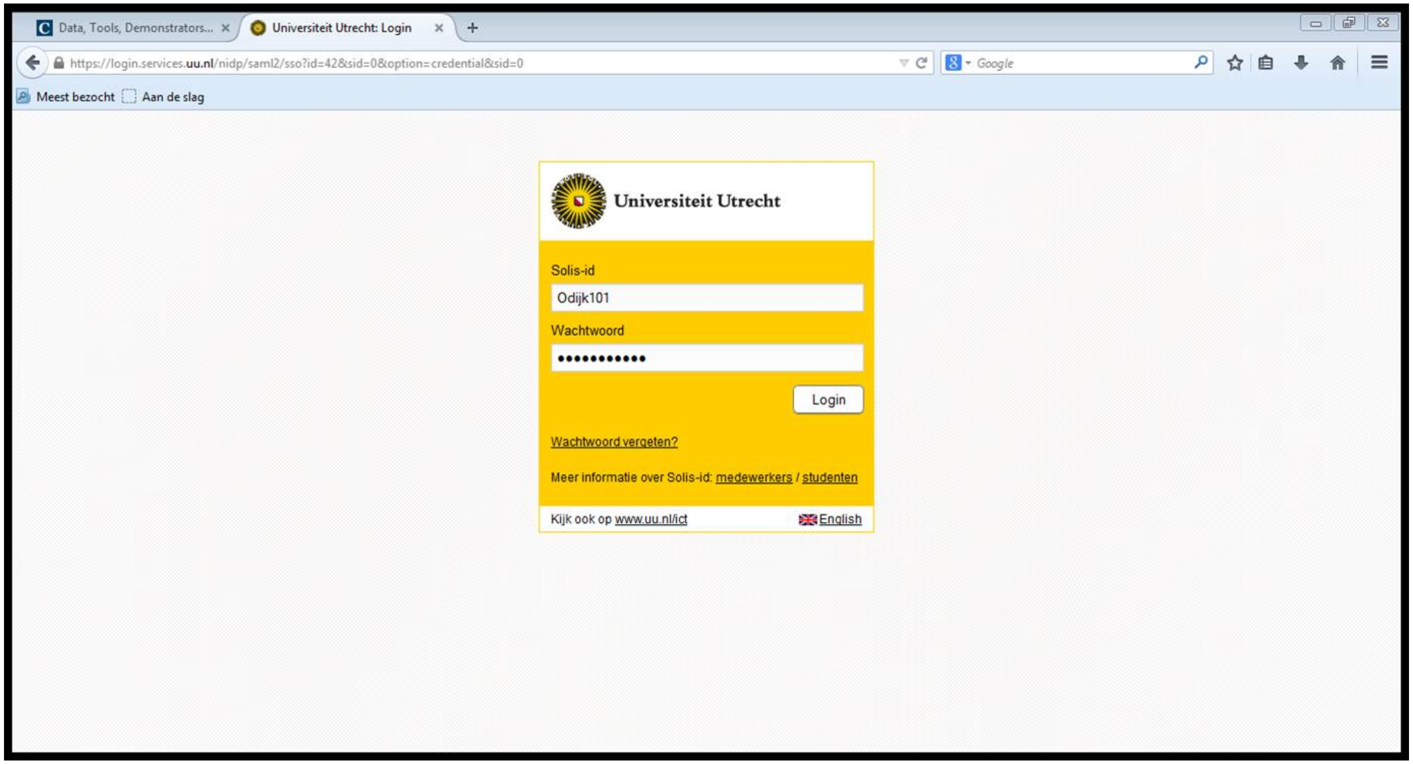
Universiteit van Amsterdam
Netherlands

Tilburg University
Netherlands

Vrije Universiteit
Netherlands

NIOO (KNAW)
Netherlands

Figure 13: You are redirected to a login via your own institute


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

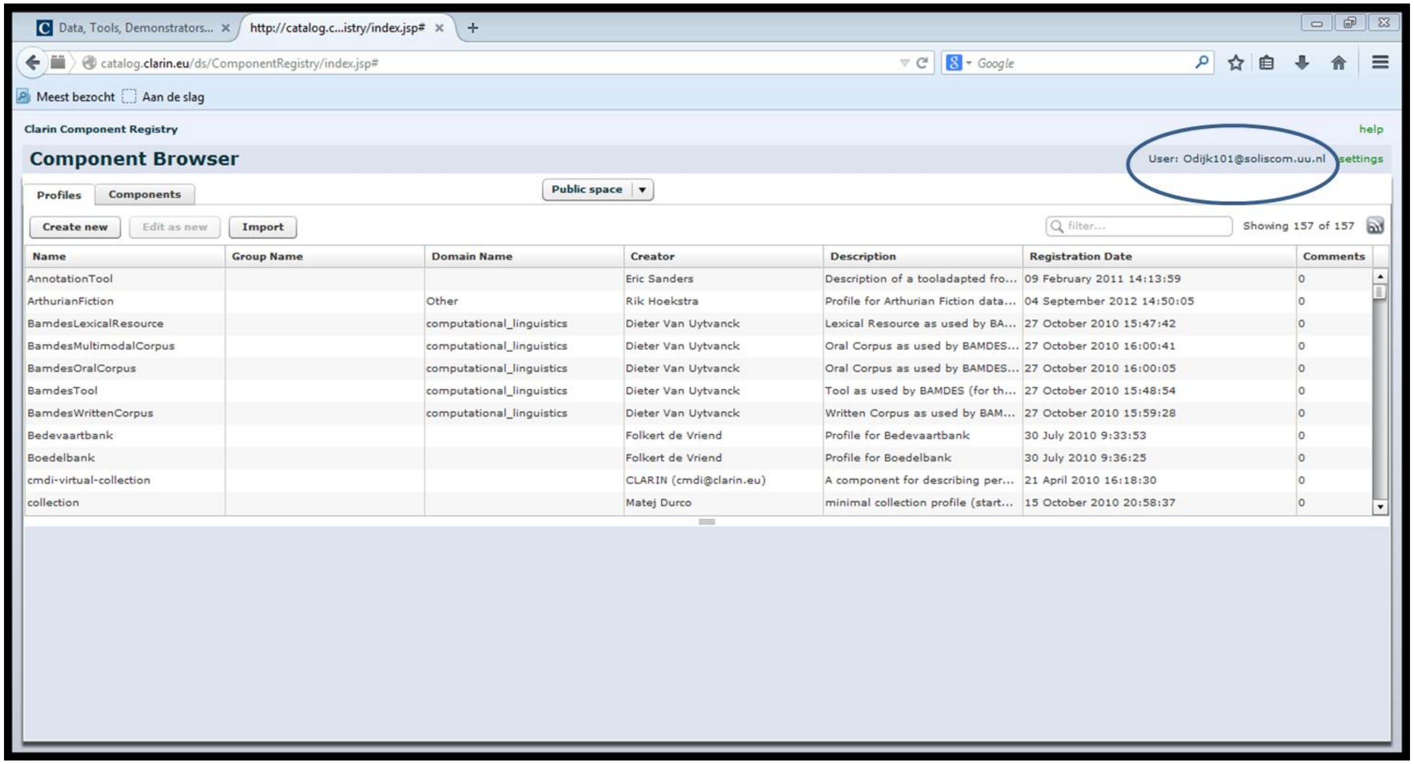
Wachtwoord vergeten?

Meer informatie over Solis-id: medewerkers / studenten

Kijk ook op www.uu.nl/ict

English

Figure 14: You login with your institute's user name and password.



Data, Tools, Demonstrators...
http://catalog.c...istry/index.jsp#
catalog.clarin.eu/ds/ComponentRegistry/index.jsp#
Google
Meest bezocht
Aan de slag
help
settings
Profiles
Components
Public space
Create new
Edit as new
Import
filter...

Clarin Component Registry

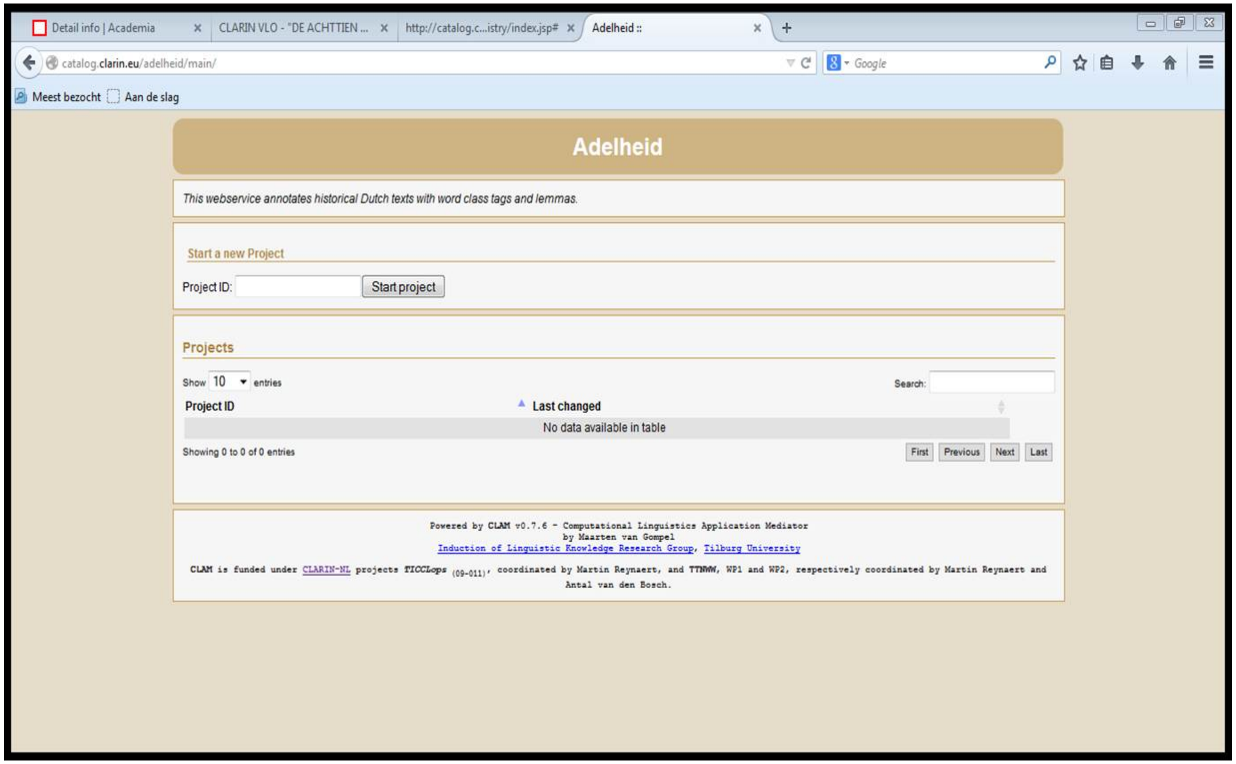
Component Browser

User: Odijk101@soliscom.uu.nl

Showing 157 of 157

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

Figure 15: You get access to the application.


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.

Figure 16: Other CLARIN services (e.g. the Adelheid application), wherever they are located, now ‘know’ that you are a trusted user, and no further login is needed.

Long Term Preservation Finally, the CLARIN centre must ensure long term preservation of your resource: it must make sure that it is still accessible after ten or twenty years or longer. CLARIN centres have to make special provisions for that. Sometimes they take care of long term preservation themselves (e.g. DANS), but most centres outsource it to specialized centres (e.g the MPI/TLA outsources it to the long term preservation services of the Max Planck Gesellschaft). In any case, each CLARIN centre must have a clear procedure in place for ensuring long term preservation, and work according to this procedure. This is one of the ingredients of the Data Seal of Approval (DSA), which each CLARIN centre must be awarded if it is to become a certified CLARIN Centre.[30] All CLARIN centres in the Netherlands have been awarded the Data Seal of Approval[31] and most are CLARIN-certified centres.[32]

6.5 Existing Resources

If you already have a resource, or have partially created it, the things that have to be done are basically the same as when you start with a new resource. However, since you already have selected a format for your resource, and possibly also for the associated resource descriptions, you probably have to adapt your resource to the requirements of CLARIN (this is called *resource curation*). Again, it is very important to contact a CLARIN centre as early as possible, because they may be able to help you with this. If the format of your resource is sufficiently formalized, it may be possible to convert it automatically to a CLARIN-compatible format. The same is true for your resource descriptions: if they are in a sufficiently formalized notation, it may be possible to convert them automatically into a CMDI format.

The CLARIN-NL project has financed many such resource curation projects. It has also set up a Data Curation Service: a team of specialists dedicated to the curation of important data for humanities researchers. In the CLARIAH successor project, such resource curation activities will be continued (as of January 2015), and you may be able to apply for funding for a project to curate your resource.

7 Concluding Remarks

I have briefly described what CLARIN is, and what it is intended for, with a strong focus on CLARIN in the Netherlands. I have described in more detail what functionality CLARIN aims to offer, and what is available at this point. Though these descriptions can serve to get a first global picture of CLARIN, additional documentation must be read and/or courses attended for really ensuring optimal use of the functionality offered. I refer to the CLARIN and CLARIN-NL web sites for additional sources, for educational and training events, and for educational packages that are being developed for use in the curriculum of humanities students.

[30] This DSA consists of 16 guidelines for the curation of data, 3 of which apply to the data producer (i.e., you!), and 3 to the data consumer (that is also you!), so it is well worth reading. The remaining 10 guidelines apply to the centre.

[31] See `http://www.datasealofapproval.org/en/community/`.

[32] See `http://www.clarin.eu/content/certified-centres`.

It must be clear from this paper that the CLARIN infrastructure already has a lot to offer to humanities researchers. However, there is also still a lot to do: many parts of CLARIN are incomplete, fragile, sometimes just prototypes instead of stable services, and for many aspects further improvements and extensions are desired or required both in terms of the functionality offered and in terms of user-friendliness. These form important challenges for the near future. In the Netherlands, the CLARIAH project, which will continue the Netherlands contributions to the design and construction of the CLARIN and DARIAH infrastructures starting in 2015, will have to take up these challenges.

Acknowledgments

This work was financed by CLARIN-NL.

References

[Broeder *et al.*, 2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valetta, Malta, 2010. European Language Resources Association (ELRA).

[Kemps-Snijders *et al.*, 2010] M. Kemps-Snijders, M.A. Windhouwer, and S.E. Wright. Principles of ISOcat, a data category registry. Presentation at the RELISH workshop Rendering endangered languages lexicons interoperable through standards harmonization Workshop on Lexicon Tools and Lexicon Standards, Nijmegen, The Netherlands, August 4-5, 2010. `http://www.mpi.nl/research/research-projects/language-archiving-technology/events/relish-workshop/program/ISOcat.pptx`, 2010.

[NWO, 2014] NWO. *NWO Subsidieregeling 1 Mei 2011 (Versie juli 2014)*. NWO, The Hague, July 2014. `http://www.nwo.nl/documents/nwo/juridisch/regeling-subsidieverlening-nwo`.

[Odijk and van Hessen, 2011] Jan Odijk and Arjan van Hessen. Sharing resources in CLARIN-NL. In *Proceedings of the Language Resources, Technology and Services in the Sharing Paradigm workshop at IJCNLP 2011*, pages 98–106, Chiang Mai, Thailand, 2011. IJCNLP 2012. `http://www.clarin.nl/sites/default/files/restore/CLARIN-NLijcnlp2011_110811.pdf`.

[Odijk, 2014a] Jan Odijk. Clarin for linguists: Concluding overview, 2014. Presentation held at the LOT Summerschool.

[Odijk, 2014b] Jan Odijk. Clarin for linguists: Introduction, 2014. Presentation held at the LOT Summerschool.

[Odijk, 2014c] Jan Odijk. Clarin for linguists: Portal and searching for data, 2014. Presentation held at the LOT Summerschool, Nijmegen.

[Odijk, 2014d] Jan Odijk. Clarin for linguists: Search illustration 1, 2014. Presentation held at the LOT Summerschool.

[Odijk, 2014e] Jan Odijk. Clarin for linguists: Storing resources in clarin, 2014. Presentation held at the LOT Summerschool.

[Odijk, 2014f] Jan Odijk. The CLARIN infrastructure in the Netherlands: Design and construction. unpublished article, Utrecht University, August 2014.

[Van Uytvanck, 2014] Dieter Van Uytvanck. How can I find resources using CLARIN? Presentation held at the *Using CLARIN for Digital Research* tutorial workshop at the *2014 Digital Humanities Conference*, Lausanne, Switzerland. `https://www.clarin.eu/sites/default/files/CLARIN-dvu-dh2014_VLO.pdf`, July 2014.

[VSNU *et al.*, 2014] VSNU, KNAW, and NWO. *Standard Evaluation Protocol 2015-2021: Protocol for Research Assessments in the Netherlands.* KNAW, Amsterdam, 2014. `https://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-2015-2021`.

[Zhang *et al.*, 2012] Junte Zhang, Marc Kemps-Snijders, and Hans Bennis. The CMDI MI search engine: Access to language resources and tools using heterogeneous metadata schemas. In P. Zaphiris et al., editor, *Proceedings of Theoretic and Practice Digital Libraries Conference (TPDL 2012)*, volume 7489, pages 492–495, Berlin / Heidelberg, 2012. Springer.