CHAPTER 2

The CLARIN infrastructure in the Low Countries

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

ABSTRACT

In this chapter I will describe what the CLARIN infrastructure is and how it can be used, with a focus on the Low Countries (and especially the Netherlands) part of the CLARIN infrastructure. I aim to explain how a Humanities researcher can use the CLARIN infrastructure. I describe the basic functionality that CLARIN aims to offer, including searching for data and software, applying software to data, and storing data and software resulting from research.

2.1 Introduction

In this chapter I will describe what the CLARIN infrastructure is and how it can be used, with a focus on the Low Countries (and especially the Netherlands) part of the CLARIN infrastructure. I aim to explain how a Humanities researcher can use the CLARIN infrastructure.[1]

The CLARIN infrastructure aims to offer 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 software resulting from the research;

and the researcher should be able to access all this functionality via a single portal.

[1] There is a series of presentations covering the major contents of this chapter. See http://www.clarin.nl/node/1959.

How to cite this book chapter:
Odijk, J. 2017. The CLARIN infrastructure in the Low Countries. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 11–30. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.2.

I will discuss each of these aspects in the sections to follow: finding data and software in section 2.2, applying software to the data in section 2.3, storing data and software in the CLARIN infrastructure in section 2.4, and the portal in section 2.5. I will end with concluding remarks (section 2.6).

2.2 Finding Data and Software

An essential function offered by CLARIN is the possibility to find resources (data and software) that might be relevant to one's 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 still there 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 will be a lot of work to select the relevant ones.

CLARIN offers this functionality of finding relevant resources as follows. First, it offers descriptions of all resources (such descriptions are known as *metadata*). Such *metadata*[2] are made in the *CMDI* format (Broeder et al., 2010). *CMDI* stands for *Component-based Metadata Infrastructure*, and it offers a flexible format for representing descriptions of resources. CMDI prescribes the format of the metadata but not their contents: these are determined by the data provider. I will go deeper into CMDI in section 2.4.

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*).[3]

Third, all metadata records are referred to via *persistent identifiers (PIDs)*, i.e identifiers that are guaranteed to exist and correctly refer persistently. The resources themselves are accessible through the metadata.

Fourth, CLARIN offers browsers and search engines to browse and search for resources via their CMDI metadata. Such browsers and search engines operate on a database of CMDI metadata located on a server of a specific CLARIN centre that acts as a metadata service provider. This database is filled and regularly updated[4] by 'metadata harvesting', i.e. an automatic process of collecting all metadata records made available by the various CLARIN centres (using the OAI-PMH protocol) and storing them in a single database.

Currently, CLARIN offers two browsers and search engines to search for resources via their metadata, viz. the *Virtual Language Observatory (VLO)*, which will be discussed in section 3.6.1, and the *Meertens CLARIN Metadata Search Engine*, which will be discussed in section 2.2.2.

Which resources can one currently find in the CLARIN infrastructure? There are several. First, there are the data and software owned by the CLARIN centres themselves (e.g. the Corpus Gysseling and associated search engine at INT). Second, there are the data and software hosted by a CLARIN 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 *CLARIN-NL Data Providers* or Type D CLARIN centres[5]), which distribute data independently of (and long before) CLARIN, but have made provisions to give access to the data that

---

[2] Though I prefer the use of the term *resource description* instead of *metadata* for the reasons sketched in Odijk and van Hessen (2011:100), I will use the term *metadata* in this book.

[3] Lagoze et al. (2002)

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

[5] This type of CLARIN centre is currently only found in the Netherlands.

are relevant to Humanities researchers in a CLARIN-compatible manner (via CMDI metadata). Examples are the National Library, the Netherlands Institute for Sound and Vision and Utrecht University Library (see chapter 4 for more details).

2.2.1 Virtual Language Observatory

The Virtual Language Observatory (VLO) offers facilities for browsing and searching in CMDI metadata. Once the desired metadata have been found, links to the actual resources (data and software) enable researchers to make use of the resources in their research.

The VLO enables a user to do a keyword (string) search for keywords that occur in the metadata. When one types in a keyword, the VLO provides suggestions for keywords that occur in the metadata (query completion): for example, if one starts typing *tree*, one gets suggestions such as *treetagger*, *trees*, and *treebank*. In addition to keyword search, the VLO offers *faceted browsing*: one can select values for a range of facets such as *language, collection, resource type, country, modality, genre, subject, format, organisation, national project, keyword* and *data provider*. For example, if one has selected *treebank* as a keyword, one can narrow down the search results to treebanks for the Dutch language by selecting *Dutch* in the *language* facet, yielding the 15 metadata records for Dutch treebanks in the VLO at the time of writing (November 2016). The VLO currently gives access to around 900K metadata records, and this number is expected to grow considerably in the coming years.[6] One can find the data dealt with in the CLARIN-NL project, as well as the data provided by the Dutch Language Union via the HLT-Agency (TST-Centrale), currently hosted by the certified CLARIN centre INT. For more information on finding data through the VLO, I refer to Van Uytvanck (2014).

2.2.2 Meertens CLARIN Metadata Search

The Meertens CLARIN Metadata Search Engine (Zhang et al., 2012) offers an alternative way to find resources through metadata. This search engine operates in principle on the same metadata as the VLO: the metadata harvested for the VLO. But snapshots from the metadata harvested for the VLO are taken at specific intervals, so there may be a difference between what is visible via the Meertens Metadata Search and the VLO.[7]

The Meertens CLARIN Metadata Search Engine also offers keyword (string) search, and it offers query completion but now on all keywords that occur in the metadata. It also indicates in which metadata element the keyword occurs and how often. This helps in selecting the desired or most relevant metadata records. 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 of how often it occurs (403 times at the time of writing) in the *description* element of the metadata element *time coverage* (see left top corner of Figure 2.1).

---

[6] The count of the number of metadata records was done in November 2016. However, this number does not say very much, because different providers of metadata may have different views on the granularity of the metadata: in some cases a metadata record 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.

[7] And in the meantime (November 2016) even these snapshots are not taken anymore, so that one finds much less data here than via the VLO.

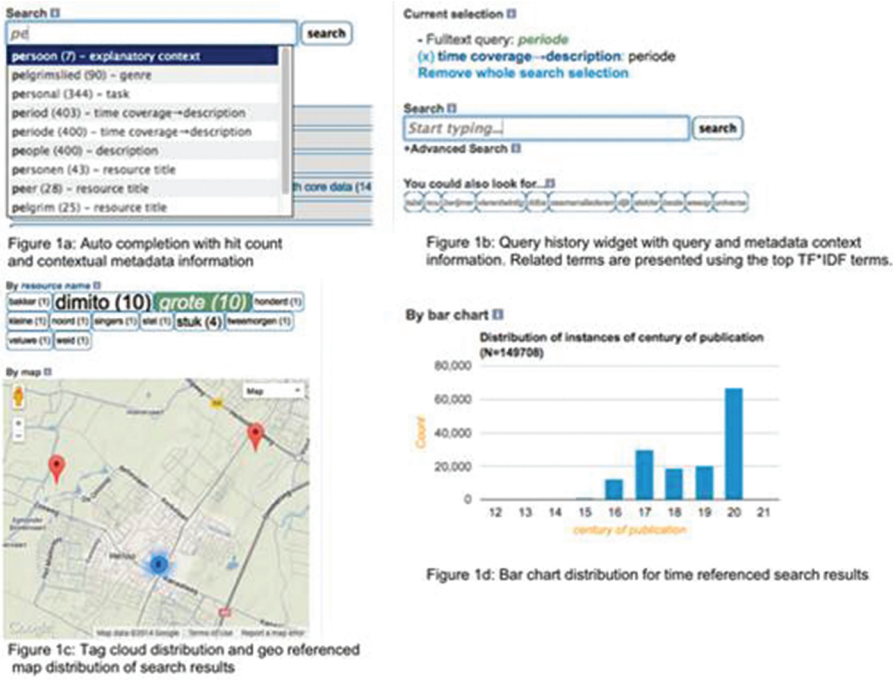

Figure 2.1: Meertens CLARIN Metadata Search Interface.

The interface also makes suggestions for other searches (see under *You could also look for...* in the mid right part of Figure 2.1). Keywords suggested there form the most important keywords related to the query based on the TF-IDF statistics.[8]

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 metadata element (see mid left part of figure 2.1), 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 2.1), and there are editable charts for displaying the date ranges of documents (see right bottom part of Figure 2.1).

Finally, it recommends related resources (see Figure 2.2) by providing links to related metadata records and a snippet of the first recommended metadata record.

2.3 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 in any detail. Instead, we will briefly describe what *types* of tools and services CLARIN currently contains, give a few concrete examples with a short description and a pointer to the CLARIN-NL portal, and mostly refer to other parts of this book where the application is described in more detail, or to other literature.

[8] A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Salton et al., 1975).

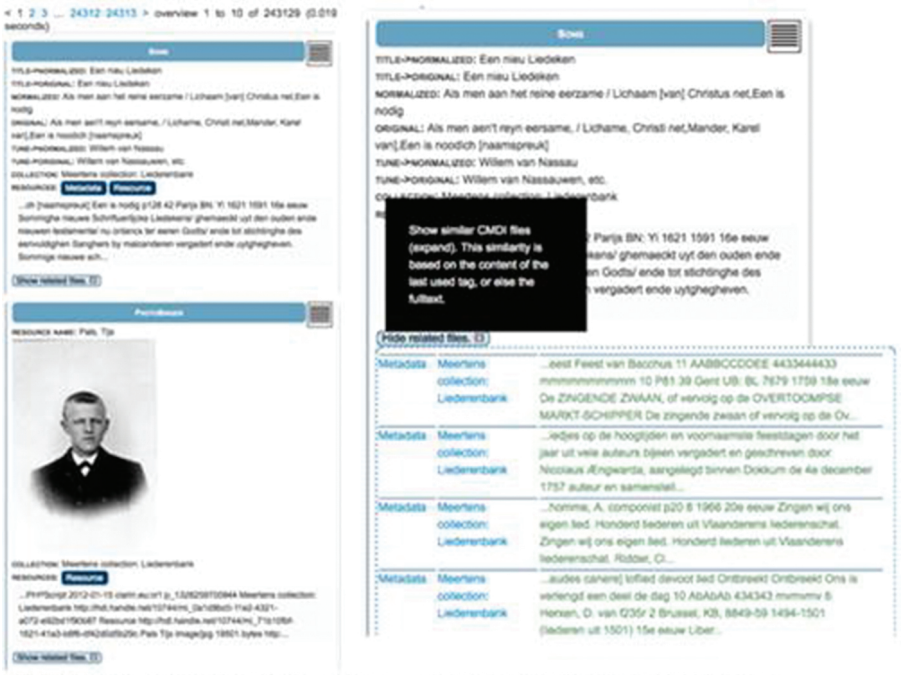

Figure 2.2: Meertens CLARIN Metadata Search Interface: recommended resources.

The tools and services can be found most easily via the CLARIN-NL portal, under Services.

Three major classes of applications and services will be discussed: searching in data (section 2.3.1), annotation and related tools (section 2.3.2), and processing data (section 2.3.3).

2.3.1 Searching in Data

Federated Content Search is a technique in which a single query can be launched to search *through* multiple resources that are stored in a different locations and that may each have their own particular format. A limited form of federated content search is possible in data via the CLARIN-D Federated Content Search graphical user interface (FCS). 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.[9]

There are also many search engines that apply to specific resources only. They include search engines for searching *in* a wide variety of resources covering a wide variety of disciplines, including literary research, historical research, religion research, media research and social research. See part IV, chapter 25 for a more detailed overview of these search applications and the other chapters in part IV for a detailed description of selected search applications.

Not surprisingly, search for linguistic properties is prominently present, e.g. through search in typological databases, in text corpora, and in lexical resources. Some applications focus on the

[9] At the time of writing (November 2016) one could search in resources from at least CLARIN-NL (though only in data at MPI), CLARIN-D, LINDAT (CLARIN-CZ) and CLARIN Poland. See `https://centres.clarin.eu/fcs` for a full overview of the current endpoints.

analysis of language variation. The scientific grammar of Dutch in the *Taalportaal* contains links to these search applications. This is described in more detail in part II (for linguistics) and in part III (for syntax).

2.3.2 Annotation and Related Tools

A number of tools focus on annotating resources, i.e. enriching them with new information. They include a web service AAM-LR for annotating where in an audio file there is speech (instead of other sounds), and identifying who is speaking in the parts containing speech (diarisation). Many improvements were made in ELAN and ANNEX, tools for the creation of complex annotations on video and audio resources, and in some closely related tools. In the SignLinC project it was made possible to link lexical databases and annotated corpora of signed language in these tools. The ColTime project extended ELAN and ANNEX with a referencing and note exchanging system. The EXILSEA project enhanced these tools for users of different languages with multilingual features. The MultiCon project enhanced ELAN and ANNEX with multilayer visualisation of multilayer collocates. TQE is a web application for evaluating the quality of phonetic transcriptions of speech files.

Several of the tools for automatic enrichment (described in section 2.3.3) can also be used for annotation purposes. They can bootstrap the annotation by automatically enriching a resource with annotations, followed by manual verification and correction. The FLAT application described in chapter 6 is an application for manual verification and correction of annotations on text corpora encoded in the FoLiA format, and the ELAN and ANNEX tools mentioned above can be used for annotating multimedia resources.

2.3.3 Processing Data

Tools for processing data include a tool for orthographic normalisation (TICClops), which is also embedded in a workflow for converting digital images into textual resources in TEI[10] format (@PhilosTEI, see chapter 32); a tool chain and methodology for converting legacy data sets in the area of maritime history (DSS); an application to analyse writing style (Stylene, see chapter 16)); and a set of web services for format conversions between a variety of formats for textual resources (OpenConvert).

It also includes tools for tokenising, lemmatising, part of speech tagging (Adelheid) and parsing (INPOLDER) of mediaeval Dutch. This functionality is also offered for modern Dutch, together with tools for assigning semantic roles and co-reference relations, and for identifying and analysing named entities. In addition, there are tools for the automatic orthographic transcription of the speech in audio files. Most of these have been implemented as web services or as workflows of web services, in particular in the TTNWW application (see chapter 7). PaQu (see chapter 23) invokes the Alpino parser to parse text corpora and makes the resulting treebank available for search and analysis.

2.4 Storing Data and Software in CLARIN

If a researcher has a resource or is going to create one, e.g. in the context of a research project, he/she can store this resource in the CLARIN infrastructure, and every researcher is strongly recommended to do so. Of course, the resource must meet the CLARIN requirements before it can enter the CLARIN infrastructure.

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

I will first discuss why it makes sense to store one's resource in CLARIN (section 2.4.1). Next, I will describe how one should store a resource in CLARIN, initially focusing on new resources. In storing a resource in CLARIN, two parties are involved: the resource provider (usually a researcher or research group that has created a resource), and a CLARIN centre. I will describe the responsibilities of the resource provider (section 2.4.3) and the responsibilities of the CLARIN centre (section 2.4.4), initially for new resources. Finally, I will discuss what has to be done for resources that already exist (section 2.4.5).

2.4.1 Why Should One Store One's Resource in CLARIN?

The first question that arises when one has a resource is: why store it in the CLARIN infrastructure?

Well, there are many reasons. I summarise them here:

Benefits to the researcher A very important reason is that the researcher may benefit from doing so: if one makes one's resource ready for storage in CLARIN, one has to put the data in a CLARIN-supported format. As a consequence, one may easily make use of existing software and data in CLARIN, so that one's data or software can be produced more efficiently, with better quality and/or with more features. One may also use CLARIN tools such as search engines, analysis tools, and visualisation tools on one's resources, so that the resource can be used immediately in research. And when one's resource is in the CLARIN infrastructure, one can be sure it is stored safely, always easily findable and accessible in ways that respect any legal or ethical restrictions, and one does not have to worry about these data in a world where software updates and upgrades are frequent so that 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 created in an earlier project in the next one. However, when one does need one's resource in a later stage, one does not have to worry where it is, and whether the medium it is stored on is still working: one can be sure to find it and get access to it via CLARIN.

Benefits to others A second reason is that others may benefit from one's resource. There are always unexpected uses of research data, immediately or only years or even decades later. CLARIN ensures that all researchers have access to the resources used in or resulting from research. Furthermore, making one's 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 these resources.[11]

Better science 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 one's research data and software is essential for that, and CLARIN provides ideal facilities for this.

Better publications Since openness about research data and results is an essential ingredient for the proper conduct of science, more and more scientific journals are beginning to require that one publishes one's research data and software, so that the results are verifiable and replicable. For the same reasons, funding agencies are also 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[12] and are available for verification and replication purposes.[13]

---

[11] Here is a clip by DANS on the importance of data sharing (in Dutch).

[12] Which, unfortunately, has happened a lot in the past.

[13] See for example, for the Netherlands, NWO (2014:19, article 30).

Benefits to the researcher's institution Increasingly, evaluation of research units includes requirements on data management and integrity. For example, 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). Each research unit wants to meet such evaluation requirements and will therefore most likely require that every researcher deals carefully with data: CLARIN offers the facilities for this.

2.4.2 How to Store Resources in CLARIN

If one's research is expected to lead to new resources, it is important to immediately start taking into account that they will be stored in the CLARIN infrastructure. Ideally, one starts with this before any data or software have been produced. If part or all of one's resources have already been produced, see section 2.4.5.

Two parties are involved in storing resources in CLARIN: the 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 2.4.3, the responsibilities of the CLARIN centre in section 2.4.4.

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

CLARIN centres come in different types.[14] The type relevant in this context is type B.

The Netherlands has multiple Type B CLARIN centres. They include the Meertens Institute (Amsterdam), the Language Archive (TLA) of the Max Planck Institute for Psycholinguistics (MPI, Nijmegen), Huygens ING Institute (The Hague), and the Institute for the Dutch Language (INT, Leiden).[15] These centres are certified CLARIN centres, which provides confidence that one's data are safely stored there in a CLARIN-compatible way. The Data Archiving and Networked Services (DANS, The Hague) is not certified as a CLARIN centre yet, but is also a reliable data centre. Which one to choose? Well, that depends on the type of resource one has and its primary intended research use. The CLARIN Portal provides information about the various centres and the types of resources they are most suited for. See chapter 4 for more details.

2.4.3 The Resource Provider

The first thing to do is to define clearly what the resource is going to be. Once this is clear, one can select a CLARIN centre, and contact this centre.[16] Next, one has to ensure that legal and ethical issues do not prevent incorporation of the resource in the CLARIN infrastructure and making it available to other researchers. There are several ways of doing this, depending on the type of resource. If the owner of the resource is a third party, the resource provider will have to obtain explicit permission for this through some licence agreement. If subjects participate in a resource creation project, one will have to ask them explicit permission to use the resource in the CLARIN infrastructure. The CLARIN centre can help with this, and there are templates for licence agreements, as well as a licence category calculator on the European CLARIN website. Together with the centre, the resource provider will have to ensure that ethical issues (mostly privacy issues), where they arise, are properly dealt with.

---

[14] This document contains an overview of the different types of CLARIN centres.

[15] Formerly the Institute for Dutch Lexicography (INL).

[16] Contact information for CLARIN centres can be found here.

We will discuss the tasks of the resource provider, initially focusing on data. We dedicate a separate paragraph to the case where one's resource is software.

CLARIN-recommended formats The resource provider has to determine a CLARIN-recommended format for the resource. A list of CLARIN-recommended formats, protocols, etc., can be found here. It is strongly recommended to consult the CLARIN centre on this issue, or to ask help from the CLARIN-NL helpdesk (`helpdesk@clarin.nl`). Since we are in the area of research, it is possible that the resource is of a completely new type, for which no CLARIN-recommended format exists. It is also possible that none of the CLARIN-recommended formats can accommodate all elements of the resource, even though the resource is not of a completely novel type. In all these cases, one has to consult the CLARIN-NL helpdesk first before continuing.

Metadata One or more descriptions must be made of one's resource. These metadata must be in CMDI-format. CMDI (Component MetaData Infrastructure) provides a model for metadata, and a format for them. It also provides tools to make metadata records. CMDI metadata are written in XML (eXtensible Markup Language). CMDI does not in any way prescribe the contents of the metadata. That is completely up to the resource provider (though CMDI helps researchers in several ways to create correct and 'useful' metadata).

CMDI metadata are structured in accordance with a *profile*. A profile describes which elements can or must be used in a metadata record. Metadata elements are XML elements, consisting of a *name*, a *value* in accordance with a *value scheme*, and a (possibly empty) set of attribute-value pairs. The definition of a CMDI element is illustrated in (1):

(1) **Element:** ResourceName
Value scheme: string
Attribute-Value Pairs

ConceptLink: `http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5`
Number of occurrences: 1 - unbounded
Multilingual: yes

It describes a metadata element called *Resourcename*, of type *string*, that must occur once but can occur multiple times. The contents can be in multiple languages. We discuss the *ConceptLink* below. Often, a group of such elements naturally belong together, e.g., because they describe a particular aspect of a resource together. One can group such elements in a metadata *component*. This enables one to treat such a collection of metadata elements as a unit. Metadata components consist of a combination of components and metadata elements. An example CMDI component is illustrated in (2):

(2) **Name: Location**
Description: Component for describing a certain location (address, region, country, continent)
Composed of:

Element: Address

Value scheme: string
Attribute-Value pairs

ConceptLink: `http://hdl.handle.net/11459/CCR_C-2528_1eaf4da1-64cc-71fc-1622-bb5bfd6e52c9`
Number of occurrences: 0 - 1
Multilingual: no

Element: Region

Value scheme: string
Attribute-Value pairs

ConceptLink: `http://hdl.handle.net/11459/CCR_C-2533_fa6e1812-e29b-3cf6-e15a-50aa34b9be68`
Number of occurrences: 0 - 1
Multilingual: no

Component: Country [0 - 1]
Component: Continent [0 - 1]

This component consists of two optional CMDI-elements (*Address* and *Region*), followed by two optional components (*Country* and *Continent*).

A profile consists of a combination of components and elements. A (partial) profile description is illustrated in (3):

(3) **Name:** LexicalResourceProfile-DLU
Description: a profile for describing a lexical resource
Components

Component: GeneralInfo-DLU [1 - 1]
Component: Access [1 - 1]
Component: Project [0 - 1]
Component: Creation [0 - 1]
Component: SubjectLanguages [1 - 1]
...

It gives the name of the profile, a short decription, and a list of components that it consists of and whether these are obligatory [1 - 1], optional [0 - 1], or iterating ([0 - unbounded] or [1 - unbounded]).

This component-based system provides high flexibility: *the resource provider* determines the contents of the descriptions for the resource by defining his/her own profiles, components, and elements. CMDI helps the resource provider with this in a variety of ways:

- A list of existing profiles and components enables one to reuse what has already been made by others: it thus saves work, and one can profit from work done by others.
- A profile and component editor [login required] enables one to create one's own profiles and components if existing profiles and components are not suited.
- Metadata editors enable one to create descriptions for resources in accordance with the selected profile in an easy and user-friendly manner. One such metadata editor is Arbil; an alternative is COMEDI (Lyse et al., 2015), developed by CLARIN Norway (CLARINO).

The flexibility offered by CMDI also has some drawbacks. One has to be aware that a major purpose of metadata is the discovery of the resources by others. It is therefore important to include information that characterises this resource and distinguishes it from other resources. It is therefore also highly recommended to use certain components that contain important metadata elements one is likely to overlook if one has to make one's 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.). One should also be aware of the fact that certain properties that are 'obvious' to one researcher are not obvious to other researchers and must therefore be included in a

proper metadata record. For example, several researchers that only work with the Dutch language omitted an indication of the language of the resource in a first version of their metadata record. The same holds for the *resource type* element, which was omitted by researchers who mainly work with text corpora. The profile name (e.g. TreebankProfile) does not itself end up in the metadata record, so any information implicitly encoded in this way (i.e., that it describes a resource of type *treebank*) must be made explicit by a metadata element. It is also important to give one's resource a name: that makes referring to it much easier. And each resource should be given an explicit version number from the start: otherwise it will become very difficult to know later which version is intended.

Reusing existing profiles and components is essential for getting better metadata, since one does not have to reinvent the wheel. It is strongly advised to follow an introductory course on CMDI before making CMDI metadata.

Explicit semantics The flexibility of CMDI has other consequences as well. In rigid metadata 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 prescribed. But with CMDI, one can choose one's own profiles, components and metadata elements, give metadata elements any name one likes, and also choose the labels for the values of these elements. But then how does another researcher or a computer program 'know' what is meant?

The flexibility offered by CMDI is possible only if the semantics of the metadata elements is made explicit . The CLARIN infrastructure must 'know' what is meant with the metadata elements, otherwise it cannot use faceted browsing in the VLO or the Meertens Metadata Search Engine.

Explicit semantics for a resource or metadata record is obtained by explicitly linking each element and its possible values in the resource or metadata record to an element of a CLARIN-recognised concept or data category registry. The most prominent registry for this purpose in CLARIN until 2014 was ISOcat (Kemps-Snijders et al., 2010). ISOcat describes data categories and their properties, such as a name and definition (in multiple languages), a unique persistent identifier, the thematic domain it belongs to, and some other properties.

ISOcat was the primary semantic interoperability registry in CLARIN, but it was not the only one. For certain types of information ISOcat is not particularly suited (e.g. for 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 is dealt with in chapter 5.

In 2014, it was decided to switch to a different system, the so-called CLARIN Concept Registry (CCR) (see chapter 4). CCR is a concept registry according to the W3C SKOS recommendation (Schuurman et al., 2016). It has not really played a big role in the CLARIN-NL project, but it will be important in the CLARIAH-CORE successor project.

The values after *Concept Link* in the CMDI element descriptions in (1) and (2) are URLs that provide the link to a concept in the CCR. The concept referred to in (1) is represented in CCR as indicated in (4):

(4) **class** Concept
status approved
prefLabel@en resource name
definition@en A short name to identify the language resource. (source: CLARIN)
notation resourceName
changeNote This concept is based on the ISOcat data category: `http://www.isocat.org/datcat/DC-2544`
inScheme Metadata

inSkosCollection Metadata
textCorpusProfile UCPH
uri `http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5`
license Creative Commons Attribution (CC BY) (use the uri above for the attribution)

In order to really use the registries and tools offered effectively, one has to attend dedicated tutorials on CMDI and semantic interoperability through CCR and CLAVAS. These have been and will be regularly organised in the Netherlands. Usually, the CLARIN centre can help researchers in creating the CMDI metadata and the explicit semantics that it requires.

Operational format v. exchange/archive format In several cases, data come in two versions: a version intended for exchange and for long term preservation (exchange/archive format), and a version that is actually used in services (operational format). 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 these 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 program will usually 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 operational format version and the exchange format version remain consistent. This requires explicit versioning, and ideally the operational format version is derived from the exchange format version in a fully automated manner. The CLARIN centres can make recommendations on how to deal with such issues.

Software The 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 community 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.[17]

It is good practice to separate the program that implements the interface from the backend software that provides the core functionality of the application. This backend may contain a single software program, but it might also contain multiple programs that work together to provide the application's functionality. These programs 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 a researcher has a desktop program, (s)he will often want to turn 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 one's desktop software into a web service using the REST protocol (van Gompel (2014); see also chapter 6 and chapter 7). 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.

[17] There may be other considerations to prefer desktop over web applications. For example, 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.

A piece of software is a resource, and therefore there must be a metadata record for each piece of software.[18] A CMDI profile for the description of software exists and is further being refined (Westerhout and Odijk, 2013).[19]

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

2.4.4 Services Offered by the CLARIN Centre

The CLARIN centre assists the resource provider with his/her tasks. The centres have experience with CMDI, with semantic interoperability, with IPR and ethical issues, and with CLARIN-supported formats and protocols, so they can advise the resource provider in such matters.

Storing the resources The CLARIN centre stores the resource provider's resource in its repository. Some centres use special software for this; e.g., LAMUS is used by MPI/The Language Archive, and the DANS EASY archiving system also offers deposition facilities that can be used by users.

Metadata harvesting The centre makes the resource available and accessible in the CLARIN infrastructure for other researchers. This is done through the metadata of the resource. The centre makes the metadata of the resource available for harvesting by others through OAI-PMH.[20] Links to the actual resource are included in the metadata, and the metadata are assigned a persistent identifier (PID, see section 2.2).

Persistent Identifiers Each centre runs or uses services for the issuing, assignment and resolution of persistent identifiers, i.e., systems that issue a persistent identifier (PID) when requested and associate it to a precise location, and that, given a PID, determine the precise location of the associated resource or metadata. See chapter 3 for more details on this.

Legal and ethical restrictions The centre makes provisions for legal and ethical restrictions, so that only persons who are allowed to get access actually get access to resources that have such restrictions. CLARIN aims to make the resources available 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 restrictions can lead to various consequences: (1) a login may be required; (2) approving special usage conditions may be required; or (3) signing a separate (paper) licence agreement may be required.

Logging in Hiding resources 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 a user that has uploaded the data to apply a service to it, as well as the data that result from this service. In such a case only this researcher (or the research team (s)he belongs to) should see and be able to manipulate these data, and this researcher does not want to be bothered by data that belong to other researchers or research groups. Logging in is an essential ingredient to achieve 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 a user. 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 the user has to log in again each time a resource happens to be located at a different centre? How can it be avoided that the user has to

[18] The term 'metadata' sounds somewhat odd for descriptions of software.

[19] Its ID in the CLARIN component registry is `clarin.eu:cr1:p_1342181139640`, but it has not been published yet.

[20] Open Archives Initiative Protocol for Metadata Harvesting (Lagoze et al., 2002).

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 Authorisation Infrastructures (AAI)*: they *authenticate* a user (determine who the user is) and *authorise* the user to do some things but not others. The AAI-system used in CLARIN is SAML-based Federated Identity Management (FIM), with Shibboleth as the most popular software implementation, and it avoids the problems mentioned above.[21]

It works as follows:

- When a user logs in (for example, to edit a CMDI component in the CLARIN Component Registry, which requires login, see Figure 2.3), the user is directed to a login with the user's own institute. See Figure 2.4.
- The user then logs in with the user's institute's user name and password. See Figure 2.5.
- If the login is successful, the institute server confirms that the user is a trusted person, and the user can enter this part of the CLARIN infrastructure. See Figure 2.6.
- If the user now goes to another part of the CLARIN infrastructure that requires login (e.g the Adelheid web application), this other part 'knows' that the user is already logged in, so the user does not have to log in again: therefore this is called *Single Sign On (SSO)*. See Figure 2.7.

Logging out is not so well-defined in this Single Sign On system. If the user has logged in to a CLARIN service, and then goes to a second one (where no login is needed because the system 'knows' that the user is logged in), the user can try to log out of the first service, but then (s)he is still logged in to the second service. So if the user now goes to the first service again, (s)he does

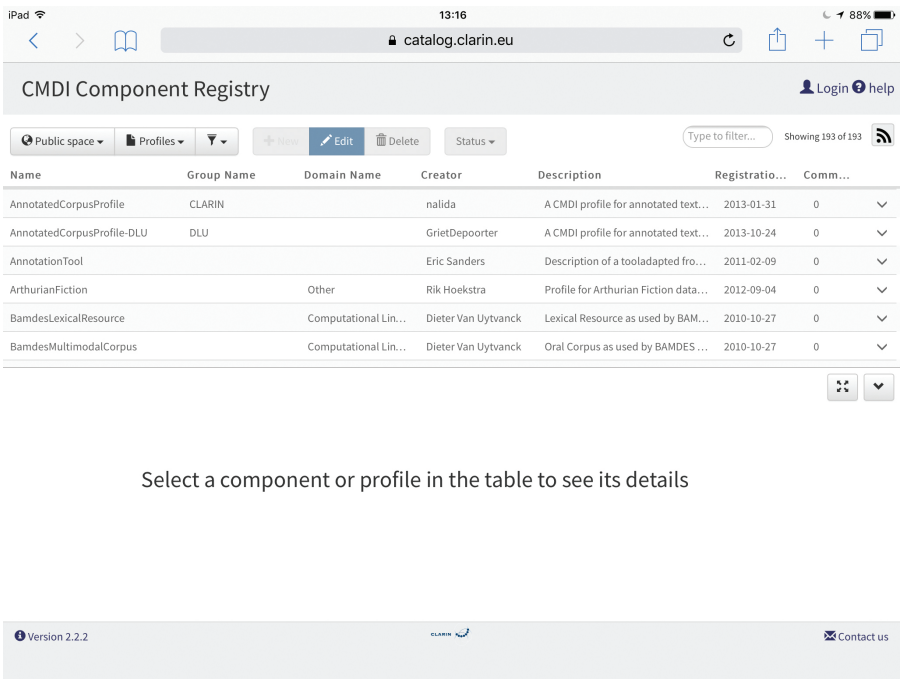

Figure 2.3: The user wants to login in the CLARIN Component Registry.

[21] CLARIN only makes use of the authentication part.

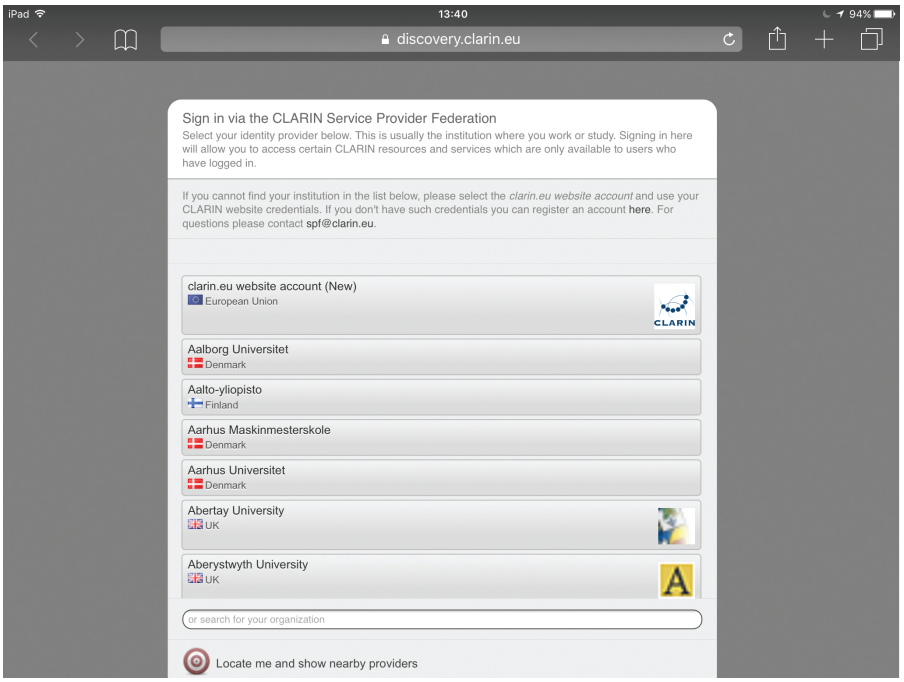

Figure 2.4: The user is redirected to a login via the user's own institute.

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) the user used to access the CLARIN services.

Long Term Preservation Finally, the CLARIN centre ensures long term preservation of the user's resource: it makes sure that it is still accessible after 10 or 20 years or longer. Centres have made special provisions in order to become certified as CLARIN centre. Sometimes they take care of long term preservation themselves (e.g., DANS), but most centres outsource it to specialised centres (e.g the MPI/TLA outsources it to the long term preservation services of the Max Planck Gesellschaft). In any case, each 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 centre must be awarded if it is to become a certified CLARIN centre.[22] All candidate CLARIN centres in the Netherlands have been awarded the Data Seal of Approval[23] and most are CLARIN-certified centres.[24]

2.4.5 Existing Resources

If a researcher already has a resource, or has partially created it, the things that have to be done are basically the same as when one starts with a new resource. However, since the researcher already has selected a format for his/her resource, and possibly also for the associated metadata, the resource probably has to be adapted to the requirements of CLARIN (this is called *resource curation*). Again,

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

[23] See `https://www.datasealofapproval.org/en/community/`.

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

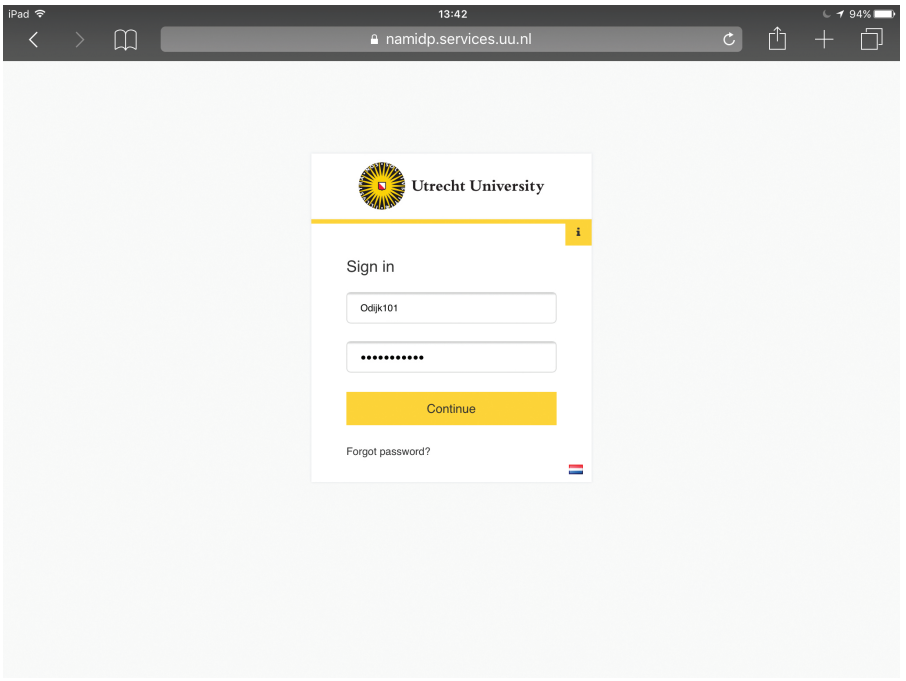

Figure 2.5: The user logs in with his/her institute's user name and password.

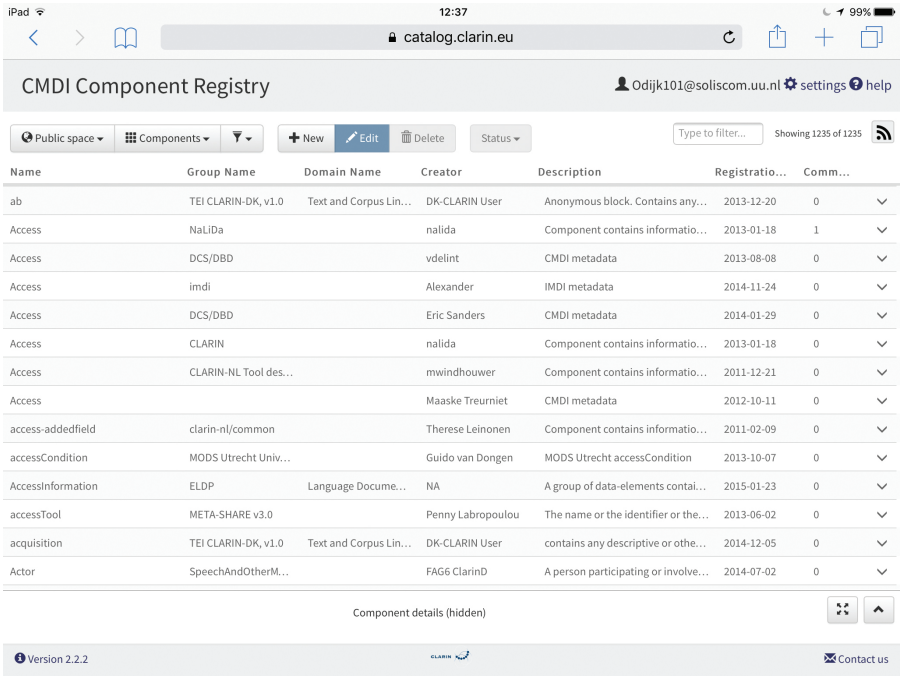

Figure 2.6: The user gets access to the application.

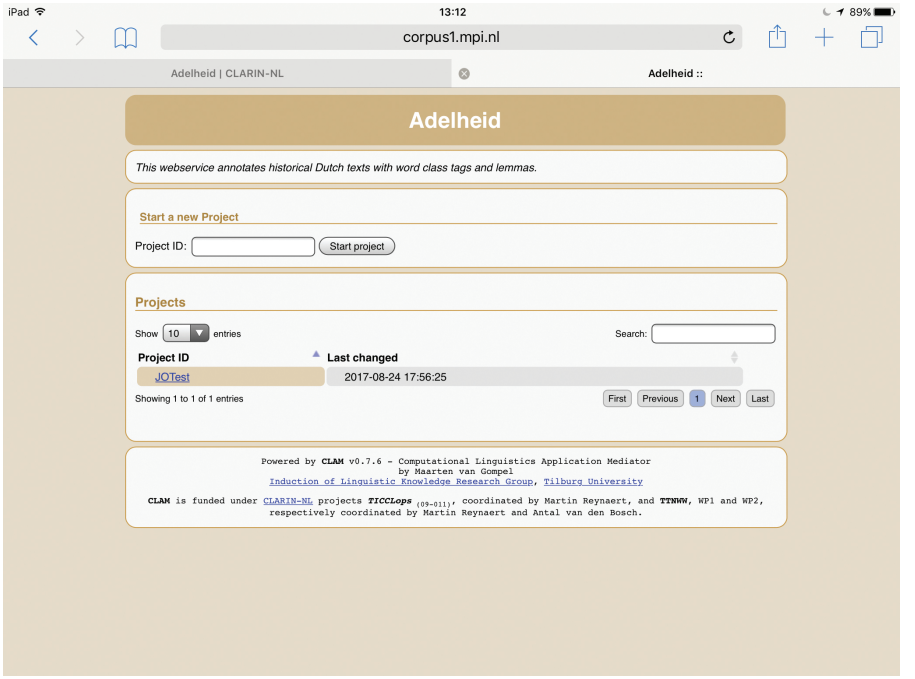

Figure 2.7: Other CLARIN services (e.g., the Adelheid application), wherever they are located, now 'know' that the researcher is a trusted user, and no further login is needed.

it is very important to contact a CLARIN centre as early as possible, because centres may be able to help with this. If the format of the resource is sufficiently formalised, it may be possible to convert it automatically into a CLARIN-compatible format. The same is true for metadata: if they are in a sufficiently formalised 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.

The curated resources include many of the data for which search and analysis applications that we mentioned earlier have been made, so these will be mentioned again in the overview given here. But they also include data that have just been curated, i.e. put into CLARIN-recommended formats, associated with CMDI metadata, where metadata are associated with PIDs, and the data stored in a CLARIN-certified centre. The types of data again cover many disciplines: within linguistics, language acquisition data, language variation data, lexical data, language documentation data, and other text corpora; for other disciplines, data for historical research, for literary research, and for religion research. They also include data from the CLARIN data providers that cover many different disciplines. See parts II, III and IV for concrete examples.

In the CLARIAH successor project, such resource curation activities have been continued, and researchers can suggest resources to be curated by the data curation service.

2.5 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, which only features a selection of everything that CLARIN has to offer, can be found via this link.

The CLARIN portal gives access to the Virtual Language Observatory (see section 3.6.1), featured resources, showcases, general information on CLARIN, CLARIN-related blogs, and instructions on how to deposit resources, and it offers the opportunity to search through multiple corpora with one query (federated search).

In addition to the Europe-wide portal, national CLARIN portals are also being created.[25] These will also 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 Low Countries can be accessed via the `http://portal.clarin.nl` URL.

This portal offers an introductory page; an overview of Dutch CLARIN centres; and a selection of tools to find relevant resources through their metadata and to search in data themselves (`http://portal.clarin.nl/node/4218`), an inventory of tools and services with faceted search on facets such as *resource type*, relevant *scientific discipline*, *tool functionality*, and others. For example, if one is interested in *syntax*, one can select that value for the facet *research discipline*; if, within syntax, one is more specifically interested in *parsing*, one 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 *TTNWW*. These descriptions also contain links to the actual services, their documentation and demonstration scenarios (see Figure 2.8). A similar faceted search interface is offered for data.

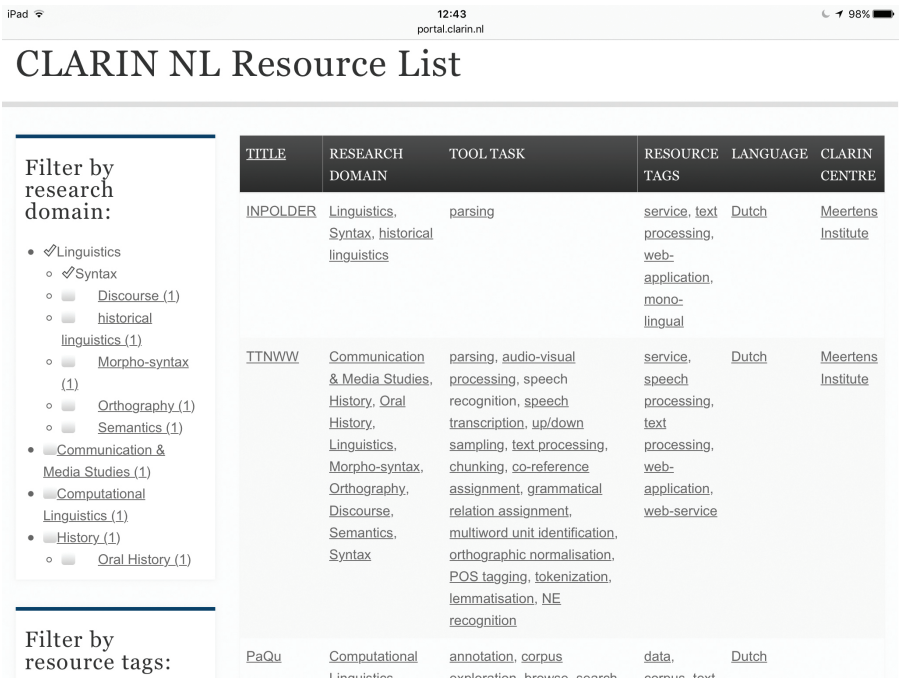

Figure 2.8: Selection of services via faceted browsing in the Dutch portal.

[25] It is not a problem that there are multiple portals, which each focuses 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.

The portal also offers a section called CLARIN recipes to get concrete guidelines in a range of matters, such as standards, issues related to intellectual property rights, how to cite data, and frequently asked questions, as well as a range of educational packages and other educational material.

2.6 Concluding Remarks

I have briefly described what functionality CLARIN aims to offer, and what is available at this point in the Low Countries. 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, CLARIN Portal and CLARIAH websites for additional sources, for educational and training events, and for educational packages that can be used in the curricula of Humanities students.

In the course of the discussion of the functionality offered by CLARIN, I have referred to many more detailed descriptions of specific functionality that will be discussed in other chapters of this book.

It must be clear from this chapter that the CLARIN infrastructure already has a lot to offer to Humanities researchers. In fact, it is already used for carrying out research, as was already pointed out in chapter 1, section 1.3. However, there is also still a lot to do: many parts of CLARIN are incomplete, fragile, and 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 continues the Netherlands' contributions to the design and construction of the CLARIN and DARIAH infrastructures starting in 2015, has taken up these challenges.

Acknowledgements

This work was financed by CLARIN-NL and CLARIAH.

References

Broeder, D., M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn (2010), A data category registry- and component-based metadata framework, *in* Calzolari, N., 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)*, European Language Resources Association (ELRA), Valetta, Malta, pp. 43–47.

Kemps-Snijders, M., M.A. Windhouwer, and S.E. Wright (2010), 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`.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner (2002), The Open Archives Initiative Protocol for Metadata Harvesting. Protocol version 2.0 of 2002-06-14, *Technical report*, Open Archives Initiative. `https://www.openarchives.org/OAI/openarchivesprotocol.html`.

Lyse, Gunn Inger, Paul Meurer, and Koenraad De Smedt (2015), COMEDI: A component metadata editor, *in* Odijk, Jan, editor, *Selected Papers from the CLARIN 2014 Conference, October 24-25*

2014, Soesterberg, the Netherlands, number 116 in *Linköping Electronic Conference Proceedings*, CLARIN, Linköping University Electronic Press, Linköping, Sweden, pp. 82–98. http://www.ep.liu.se/ecp/article.asp?issue=116&article=008&volume=.

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

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

Salton, G., A. Wong, and C. S. Yang (1975), A vector space model for automatic indexing, *Commun. ACM* **18** (11), pp. 613–620, ACM, New York, NY, USA. http://doi.acm.org/10.1145/361219.361220.

Schuurman, Ineke, Menzo Windhouwer, Oddrun Ohren, and Daniel Zeman (2016), CLARIN Concept Registry: The New Semantic Registry, *in* Smedt, Koenraad De, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14{16, 2015, Wroclaw, Poland*, number 123 in *Linköping Electronic Conference Proceedings*, CLARIN, Linköping University Electronic Press, Linköping, Sweden, pp. 62–70. http://www.ep.liu.se/ecp/article.asp?issue=123&article=004.

van Gompel, Maarten (2014), CLAM: Computational Linguistics Application Mediator. Documentation.version 0.9.12 - revision 1.1, *Language and Speech Technology Technical Report Series LST-14-02*, Radboud Centre for Language Studies, Radboud University Nijmegen, Nijmegen. http://www.clarin.nl/sites/default/files/clam_manual_2.pdf.

Van Uytvanck, Dieter (2014), 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.

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

Westerhout, Eline and Jan Odijk (2013), Metadata for tools: creating a CMDI profile for tools, Presentation held at CLIN 2013, Enschede, the Netherlands. http://www.clarin.nl/sites/default/files/13CLIN.pdf.

Zhang, Junte, Marc Kemps-Snijders, and Hans Bennis (2012), The CMDI MI search engine: Access to language resources and tools using heterogeneous metadata schemas, *in* Zaphiris, P. et al., editor, *Proceedings of Theoretic and Practice Digital Libraries Conference (TPDL 2012)*, Vol. 7489, Springer, Berlin / Heidelberg, pp. 492–495. http://www.clarin.nl/system/files/Zhang_tpdl2012.pdf.