The CLARIN infrastructure in the Netherlands: Design and Construction

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

1 Introduction

In this paper I will describe the design and construction of the CLARIN infrastructure, with a focus on the Netherlands part of it. The targeted audience of this paper is humanities researchers: I will try to explain what had to be done behind the scenes to make the CLARIN infrastructure in the Netherlands work. This paper presupposes knowledge of Odijk (2014).

CLARIN-NL is probably best known among humanities researchers in the Netherlands for the data curation and demonstrator projects[1], in which humanities researchers, in close cooperation with computer scientists and CLARIN centres, adapt their research resources to the requirements of CLARIN (*curation*), and/or create user friendly applications to browse, search or analyze research data (*demonstrators*). The focus of this article will be on aspects of the CLARIN-NL project which have been less visible to the outside world but have been crucial for a working CLARIN infrastructure. The relevant work has been carried out in subprojects that focused on the design and construction of the CLARIN infrastructure in the Netherlands, and they include the Metadata project, IIP, S&D, CLAVAS, MD4T, CLAPOP, the Data Provider subprojects, and TTNWW.

I briefly characterize the activities of these projects:

Metadata Project has developed metadata profiles and components for frequently occurring data types

IIP (*Infrastructure Implementation Project*)has implemented basic functionality of the CLARIN infrastructure in the Dutch CLARIN centres

S&D (*Search & Develop*) has developed a metadata search engine and worked on federated content search

CLAVAS (*CLARIN Vocabulary Access*) has developed a single access points to different vocabularies and concept registries

MD4T (*Metadata for Tools*) has developed a CMDI profile for the description of software

CLAPOP (*CLARIN Portal Project*) has designed and is constructing the Netherlands CLARIN portal

[1] `http://www.clarin.nl/node/281`.

Data Provider subprojects In these projects, an organisation that, by its very nature makes available a large amount of digital data, ensured that these data are available in a CLARIN-compatible manner.

TTNWW developed a work flow system for web services for language and speech technology software

See `http://www.clarin.nl/node/281` for more details on these projects. In this paper we will regularly refer to these projects when discussing specific functionality. I will also introduce some new (small) projects that have recently started or are about to start.

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 globally describe the CLARIN Infrastructure and what it is intended for (section 2). Next, I describe the CLARIN Centres in the Netherlands, on which the CLARIN Infrastructure has been built (section 3). After that, I will describe the major functionality CLARIN aims to offer to researchers: finding data and software (described in section 4), applying software to data (described in section 7), storing data and software in the CLARIN infrastructure (described in section 5, and all of that via single portal (described in section 6). I end with concluding remarks (section 8).

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.[2]

Infrastructure refers to (usually large-scale) basic physical and organizational resources, structures and services needed for the operation of a society or enterprise.[3]

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.

[2] A more detailed explanation can be found in Odijk (2014).

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

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]

Digital language resources include both data and software. As for data, it covers 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 not one big physical installation on a single location like famous research infrastructures such as the CERN Large Hadron Collider or the Chile Large Telescope. On the contrary,

- 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 3.
- 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.[5]

The CLARIN infrastructure is still under construction, is highly incomplete, and 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

[4] CLARIN at the European scale is intended for the humanities *and* the social sciences, but the Netherlands has focused on the humanities.

[5] 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.

I will discuss each of these aspects in the sections to follow: finding data and software in section 4, storing data and software in the CLARIN infrastructure in section 5, the portal in section 6, and applying software to the data in section 7. I end with concluding remarks (section 8). But I begin with a description of the CLARIN Centres in the Netherlands, because this is what the CLARIN infrastructure has been built upon (section refcentres).

3 CLARIN Centres in the Netherlands

CLARIN Centres come in different types.[6] For the Netherlands, 3 types are relevant: type B, type A, and type D CLARIN Centres.

3.1 Type B Centres

Type B Centres offer online services and harvestable metadata that are accessible in a CLARIN-compatible manner and they provide fully integrated CLARIN-conformant services.

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). They differ in the kind of resources that they are interested in, usually as a function of their research interests. 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 related to 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

Most centres provide not only data but also services.[7] DANS, however, only provides data.

[6] This document contains an overview of the different type of CLARIN Centres.
[7] Such centres are sometimes called B+ Centres, but this is not an official term.

3.2 Type A Centres

Type A Centres offer core, essential infrastructure services. For example, the MPI / TLA hosts the Virtual Language Observatory (VLO, see section 4.5.1), and carries out the harvesting of resource descriptions required for that (see section 4.2).[8]

Only the MPI / TLA and the Meertens Institute currently offer type A services (in addition to the type B services they offer). MPI / TLA offers many type A services, and Meertens a few. For example, Meertens hosts the *Meertens Metadata Search Engine* (see section 4.5.2), which has a similar function as the VLO.

3.3 Type D Centres

Type D Centres (also called *Data Providers*) 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, via CLARIN-compatible login if login is required). This type of CLARIN Centre is at this moment restricted to the Netherlands and includes organisations that, by their very mission, make available large amounts of data, currently:

Koninklijke Bibliotheek (KB) [9] Digital books, articles, newspapers. Includes the DBNL (Digital Library for Dutch Literature)

Nederlands Instituut voor Beeld & Geluid (NIBG) [10] Audio-visual data (esp. TV and radio programmes)

Utrechtse Universiteitsbibliotheek (UBU) [11] Digital books and articles

4 Find Resources

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 filter them.

CLARIN offers this functionality as follows. First, it offers, for all resources, descriptions of the resources (also known as *metadata*).[12] This will be discussed in section 4.1. Second, the resources and their resource descriptions are stored on servers of CLARIN Centres. See

[8] See `http://www.clarin.eu/content/services` for more examples.
[9] National Library.
[10] Netherlands Institute for Sound and Vision.
[11] Utrecht University Library.
[12] 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).

section 5. Third, the resource descriptions are made available to the outside world. This will be discussed in section 4.2 Fourth, all resource descriptions and all resources are referred to via *persistent identifiers (PIDs)*, i.e identifiers that are guaranteed to exist and correctly refer persistently. This will be discussed in section 4.3). Fifth, CLARIN offers browsers and search engines to browse and search for resources via their resource descriptions. These will be discussed in section 4.5.

4.1 Resource descriptions

For *Resource descriptions* both their form and their meaning must be defined. I describe the form of resource descriptions in section 4.1.1, and their meaning in the section 4.1.2. Both sections will also contain descriptions of tools that support the creation and use of resource descriptions.

4.1.1 Form

For each resource, one (or more) resource descriptions must be made. These resource descriptions (also called 'metadata') must be in CMDI-format. CMDI (*Component-based Metadata Infrastructure*, Broeder *et al.* (2010)) 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 the resource provider (though CMDI assists resource providers in several ways to create correct and 'useful' resource descriptions.

CMDI resource descriptions use 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. 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 makes it possible to treat such a collection of resource description elements as a unit. Resource description components contain resource description elements, and can contain components themselves. A profile consists of a combination of components and elements.

This component-based system provides high flexibility: the resource provider can determine the contents of the descriptions for his/her resource by defining his/her own profiles, components, and elements.

It must be possible to create profiles, components and resource description elements in a user friendly way. To that end, the design and the construction of a *profile and component editor* was started up in CLARIN-PP. The editor can be found here (login is required).

The resulting profiles, components, and elements must be stored somewhere and be accessible to other researchers and to programmes. To that end, the design and construction of the Component Registry was initiated by CLARIN-PP. It makes it possible to inspect existing profiles, components, and elements, and to reuse them for other resource descriptions.

The CLARIN-NL *Metadata Project* contributed to the construction of the profile and component editor and to the Component Registry in the early phase of the CLARIN-NL

project.

It must be possible to make new resource descriptions or adapt existing ones, using the profiles and components defined in the Component Registry. To that end, an existing resource description editor, ARBIL, was adapted so that it can work with the profiles and components defined in the Component Registry. The IIP project contributed to this adaptation.

The flexibility offered by CMDI offers many advantages but also has some drawbacks. Flexibility is needed where there are good reasons to deviate from what others have done, but may be a burden for cases where there are deviations because of lack of knowledge of what has been done before. It is therefore essential that a Component Registry exists so that reuse of profiles and components can be maximized, and unnecessary errors or omissions can be avoided. It also provide researchers with the opportunity to inspect resource profiles, which may make them aware of properties that may be 'obvious' to them but not to the whole CLARIN research community.[13]

The Component Registry has been created and is in use, but it quickly became clear that it was not easy to find components and profiles that could be relevant to your resource, since the registry consists, in essence, of a flat list of profiles and components, and search facilities are lacking. As a consequence, new users started creating their own profiles and components, which actually increased the problem of finding potentially relevant profiles and components. Also, the lack of a clear versioning strategy increased the problem.[14] A clean up action is desperately needed!

In 2014, a project has indeed been started up to investigate how the quality of existing and new resource descriptions can be improved, how reuse of existing profiles and components can be increased, and how profiles relate to one another. It is expected that an approach to address the problem will result from this project by the end of 2014. The problem is not unique to the Netherlands. Austria has encountered the problem as well, and has developed a tool, the SMC-browser, to investigate the relations between profiles and components (Ďurčo and Windhouwer (2014)).

In order to make sure that profiles for frequently occurring resource types were available before a large set of data curation projects were in need of them, early in the CLARIN-NL project the *Metadata* subproject created profiles for text corpora, lexical resources speech corpora and for a number of specific other resources in the Netherlands. This was done only for data, but not for software. Resource descriptions for software have not been made in CLARIN-NL (with 2 exceptions), but in 2011 a Metadata for Tools (MD4T) subproject was started up to create a profile for software. This profile was developed by testing it against 5 pieces of software curated in CLARIN-NL. Currently it is being refined and applied to all software curated or created in CLARIN-NL.

[13]For example, researchers only working with data from the Dutch language will not easily add a resource description element to mark the language of the resource explicitly. Inspecting other profiles will make them aware of the need to explicitly mark such information. Unfortunately, despite this, several profiles have been created where this information is lacking.

[14]For example, at a certain point there were 3 different components called *GeneralInfo* created by user *nalida*, and it was totally unclear how they were related. Currently there are still two, and their mutual relation is unclear.

4.1.2 Semantics

The flexibility of CMDI has other consequences as well. In rigid resource description schemes, the position of an element determines its interpretation (e.g in a CSV format) and (e.g. in Dublin Core) 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 person, or a computer programme 'know' what you mean with it?

The flexibility offered by CMDI requires explicit semantics! The CLARIN infrastructure must 'know what is meant with specific 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)). Its design and construction was initiated in combination with the ISO initiated ISO TC37 (Terminology and Other Language and Content Resources) technical committee, but a large part of the construction, and the main part of the maintenance since 2009 was carried out in the context of the CLARIN-NL *IIP* project.

ISOCAT offers an registry for data categories in accordance with the ISO 12620 standard, a web application for browsing, searching and editing, and a web service for communication with other programmes.

ISOCAT is basically just a flat list of data categories. There is a little bit more structure, because DCs are grouped by thematic domain, but finding an existing DC that might be relevant is quite difficult. For example, if one searches for a data category for 'grammatical relation', one will not find one with this name. Perhaps one will find the data category grammatical function because of its alphabetical closeness, but how is one ever going to find that ISOCAT also contains a data category for syntacticFunction (Odijk (2009:12))? Because of this, reuse of data categories has been minimal, and ISOCAT has seen a proliferation of near-identical data categories. This is one of the reasons why it is desirable to be able to specify relations between data categories. Relations between data categories can be used to group data categories by various criteria, which will make searching for related data categories easier, and will make it possible to consider different categories (such as *grammatical function* and *syntacticFunction*[15] as identical or near-identical. This can be done in a special registry, called RELCAT (which currently only exist in an α-version).

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. For example, the *de facto* standard for PoS-tags for Dutch (Van Eynde (2004)) is well-structured in accordance with a clearly defined syntax, which, however, is specific for this tag set. For example, a tag for nouns takes the form:

- tag = 'N ', '(', NTYPE, ',', GETAL, ',', GRAAD, ',', GENUS, ',', NAAMVAL,')

where the upper case labels between the brackets are non-terminals (corresponding to attributes and/or the types of the possible values of an attribute) that can be rewritten into

[15] and the category with name syntactic function.

terminals corresponding to the values of the attributes.

Since the syntax of such tags is idiosyncratic, standard programs (that expect e.g. XML syntax) will consider such tags as unanalyzable values. But we want to associate parts of these tags to ISOCAT DCs, e.g. the attribute *NTYPE* to `http://www.isocat.org/datcat/DC-4908`, and the value *soortnaam* to `http://www.isocat.org/datcat/DC-4910`. SCHEMACAT makes the syntax of these tags explicit so that ISOCAT DCs can be assigned to parts of the tag.

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 SIL for ISO). In order to use such other registries in a transparent manner, the CLAVAS Vocabulary Service has been set up (by the CLAVAS project) as an interface to other data category registries and vocabularies, and as a service to store data categories not dealt with elsewhere.[16]

The CLARIN Vocabulary Service is using vocabularies formulated in the Simple Knowledge Organization System (SKOS). It is a specific instantiation of the OpenSKOS exchange and publication platform for vocabularies. OpenSKOS offers multiple ways to obtain and publish SKOS vocabularies and to use vocabularies. For example, it offers search and autocompletion of vocabulary items, which can be used by other programmes (e.g. a resource description editor). CLAVAS is hosted by the Meertens Institute.[17]

CLAVAS currently publishes SKOS versions of 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[18], based on the CLARIN VLO. The curation of the names has been carried out by the CLARIN-NL Data Curation Service.

4.2 Harvesting resource descriptions

The resource descriptions must be made public, otherwise nobody knows of its existence and the existence of the resource it describes. Each CLARIN centre makes its resource descriptions available on a server that is publicly accessible. They are made available through a specific protocol called OAI-PMH[19], which allows other programs to easily access them. After all, we want to be able to find all resource descriptions via one interface. In order to make this possible, some type A centres regularly gather the resource descriptions of all CLARIN centres in one central database. This process is called resource description *harvesting*, and it is done through the OAI-PMH protocol. This has to be done regularly[20],

[16] In the Semantic Web, vocabularies define the concepts and relationships (also referred to as terms) used to describe and represent an area of concern.

[17] Another example of the type A services offered by Meertens Institute.

[18] requires login

[19] Open Archives Initiative Protocol for Metadata Harvesting.

[20] See `http://www.clarin.eu/faq/when-metadata-vlo-harvested` for the harvesting update schedule for the *Virtual Language Observatory* (see section 4.5.1).

since new resource descriptions will regularly appear at each CLARIN centre.[21]

The 'harvesting' software run by a type A centre must 'know' where the resource descriptions of each centre can be found. This is one of the reasons why the CLARIN centre registry has been set up.[22] A registry is a central database that enables one to store and maintain information, and it provides facilities to extract information from it. The centre registry has an entry for each CLARIN centre with information about this CLARIN centre (inter alia, the server where the resource descriptions are made available through OAI-PMH, its *OAI-PMH end points*).

The centre registry has been developed by CLARIN-D, and each Dutch CLARIN centre has entered the required information about itself there. Here are views on the centre registry, and here's an overview of the OAI-PMH end points.

4.3 PIDs

Locations on the internet are usually specified by means of a Universal Resource Locator (URL), such as `http://www.clarin.nl`. It is well-known that URLs often simply disappear, or change their name. This happens because the URLs are usually created and maintained by a particular project (which is temporary by nature), or by a particular organisation (which tends to be more stable but nevertheless is not immune to changes). URLs often also reflect the internal structure of an organisation, and that is surely less stable than the organisation itself.

In CLARIN we need a way to refer to objects on the internet that is more stable than using URLs. *Persistent identifiers* offer that functionality. A *persistent identifier (PID)* itself is no more than an identifier, and does not bring very much by itself. A crucial ingredient for persistent identifiers to serve their role is (1) an organisation that holds itself responsible for the PIDs it assigns, and (2) a software system that supports this organisation in the creation, the assignment, the maintenance and the resolution of PIDs.

A persistent identifier is an identifier, ideally without any internal structure or semantics, which is created by an organisation that considers itself responsible for it. A newly created PID must be unique. A PID is associated with a URL, and this relation is stored in a PID resolution system. The PID will never change. Of course, the URL it is associated with may change, or disappear, but it is the responsibility of the organisation that issued the PID to ensure that the PID will continue to refer to the same object through some other URL. Of course, an organisation can ensure this only for URLs that it controls itself.

Each resource and each resource description is assigned a PID in the CLARIN infrastructure. In this way, a user or software programme that wants to use a specific resource can simply refer to the PID assigned to it and never has to change this reference anymore. The PID resolution system will resolve the PID, i.e. replace it by the URL it is associated with, and transfer the user or the software to object(s) at this location.

Each Netherlands CLARIN centre has set up a PID-system for the creation, the assignment, the maintenance and the resolution of persistent identifiers. In CLARIN, the so-called

[21] Currently, only the MPI / TLA does such regular harvesting. Meertens Institute regularly takes a snapshot of the resource descriptions harvested by MPI / TLA for its CLARIN search engine.

[22] See `http://www.clarin.eu/blog/central-role-centre-registry` for other reasons why the centre registry is important in the CLARIN infrastructure.

Handle system is used for the assignment and resolution of persistent identifiers.[23] An example of a Handle PID is *10032/12824827a77b9602cc66840a62aedf43*. The uniqueness of each PID is guaranteed because each centre has its own prefix (10032 in this example), and the PID-system guarantees the assignment of a unique new identifier within the system. Having these PIDs preceded by the prefix `http://hdl.handle.net/` turns them into *actionable* PIDs, i.e, PIDs that are resolved and lead you to the resource it is associated to by clicking on it, just as any URL.

It is not really necessary for each CLARIN centre to run its own Handle Server installation: they might also make use of a Handle Server installation made available by external parties. In the European context, the *European Persistent Identifier Consortium* (EPIC) consortium would be a natural candidate for such an external party, but its services became available rather late and were never officially offered despite an invitation from CLARIN-NL (and their services were rumoured to be quite expensive).[24] A commercial alternative was offered by Picturae, but it became available only after each centre had already set up its own server, which made it unattractive to select this option.

4.4 Resources

In the preceding sections I described the systems used in CLARIN for resource descriptions. But which resources are in the CLARIN infrastructure? There are of course too many to list them here, even if we restrict attention to the Netherlands, but we can point out some major classes.

The CLARIN centres have made the resources curated in the resource curation and demonstrator projects available in the CLARIN infrastructure. They originate from a researcher from another research organisation. An (arbitrary) example is formed by the FESLI data and search engine. The FESLI data originate from Amsterdam University.

The CLARIN-centres have also made a part of their own resources available in CLARIN, e.g. the Corpus Gysseling and search engine, and the Corpus of Contemporary Dutch[25] at INL, the Meertens Soundbites Collection at Meertens Institute, and the DoBes Archive at MPI / TLA. Huygens has adapted its eLaborate software for scanning, transcribing and annotating textual resources, so that it can generate resource descriptions in CMDI format.

CLARIN Type D centres have many data that are highly relevant to humanities researchers, and for this reason we want these data to be available via the CLARIN infrastructure. And they are, for NIBG, for UBU, or will soon be (KB).

Unfortunately, many important resources that are relevant to the Dutch language are currently still absent, inter alia the resources for the Dutch language at the HLT-Agency (TST-Centrale). After its split from INL, activities for CLARIN were delayed. It now aims to become a certified CLARIN Centre for the CLARIN ERIC partner Dutch Language Union.

[23] DANS currently uses a different system based on *Uniform Resource Names* (URNs). It will start using Handle as well in 2014.Though the URN-based system originates from the publication-oriented library domain, Utrecht University Library uses Handle PIDs in the CLARIN context. KB*?

[24] Several CLARIN centres outside of the Netherlands make use of the EPIC services, though. See the Centre Registry.

[25] See the INL CLARIN Portal.

4.5 Browser and Search engines

With resources and CMDI resource descriptions for them, both of which can be referred to via PIDs, and with harvesting facilities through OAI-PMH, everything is in place to create functionality for browsing in and searching for resources.

This functionality is typical for CLARIN Type A centres. It requires a browsing and/or search engine in combination with a web interface. Such engines operate on a database of CMDI resource descriptions located on a server of a CLARIN Type A centre, which is filled and regularly updated by resource description harvesting, as described in section 4.2.

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 4.5.1, and the *Meertens CLARIN Metadata Search Engine*, which will be discussed in section 4.5.2.

Services offered by CLARIN Centres in the Netherlands can also be found by faceted search in the CLARIN-NL portal. This will be discussed in section 6.

4.5.1 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 allow the researcher to make use of the resources in his/her research. For certain data, there are not only links to the data, but also to software for browsing, searching and/or analysing the data.[26]

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).[27] 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 years.[28] For more information on finding data through the VLO, I refer to Van Uytvanck (2014).

The VLO has been created in CLARIN-PP, but CLARIN-NL has contributed to its maintenance and further upgrades through the CLARIN-NL IPP project. Also CLARIN-D has made major contributions to the re-design of the VLO and the implementation of upgrades.

[26] And for certain data with IPR restrictions there are only links to such software.

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

[28] 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.

4.5.2 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 at 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. In this case it does so on all keywords that occur in the resource descriptions, and it 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 1).

The interface also makes suggestions for other searches (see under *You could also look for...* in the mid right part of figure 1. Keywords suggested there form the most important keywords related to the query based on the TF-IDF statistics.[29]

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 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 1), and there are editable charts for displaying the date ranges of documents (see right bottom part of figure 1).

Finally, it recommends related resources (see figure 2) by providing links to related resource descriptions and a snippet of the first recommended resource description.

The Meertens CLARIN Metadata Search Engine has been created in the CLARIN-NL project *Search & Develop*.

5 Store data and software

CLARIN-centres must store data and software. To that end, each CLARIN centre has set up a repository.[30] Different software systems for repositories exist, and CLARIN does not proscribe which system has to be used. Common repository software systems are the open source repository platform Fedora[31], DSpace, and LAT (developed by MPI, Broeder (2009))

The Centre registry contains information on the repository systems used by the various CLARIN Centres. One can see there that the each Dutch CLARIN centre uses a different

[29] 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`.

[30] This was done in the IIP project for all centres except Huygens ING. Huygens ING joined later as a CLARIN centre and did this in a project specific for Huygens ING.

[31] *Fedora* stands for *Flexible Extensible Digital Object Repository Architecture*

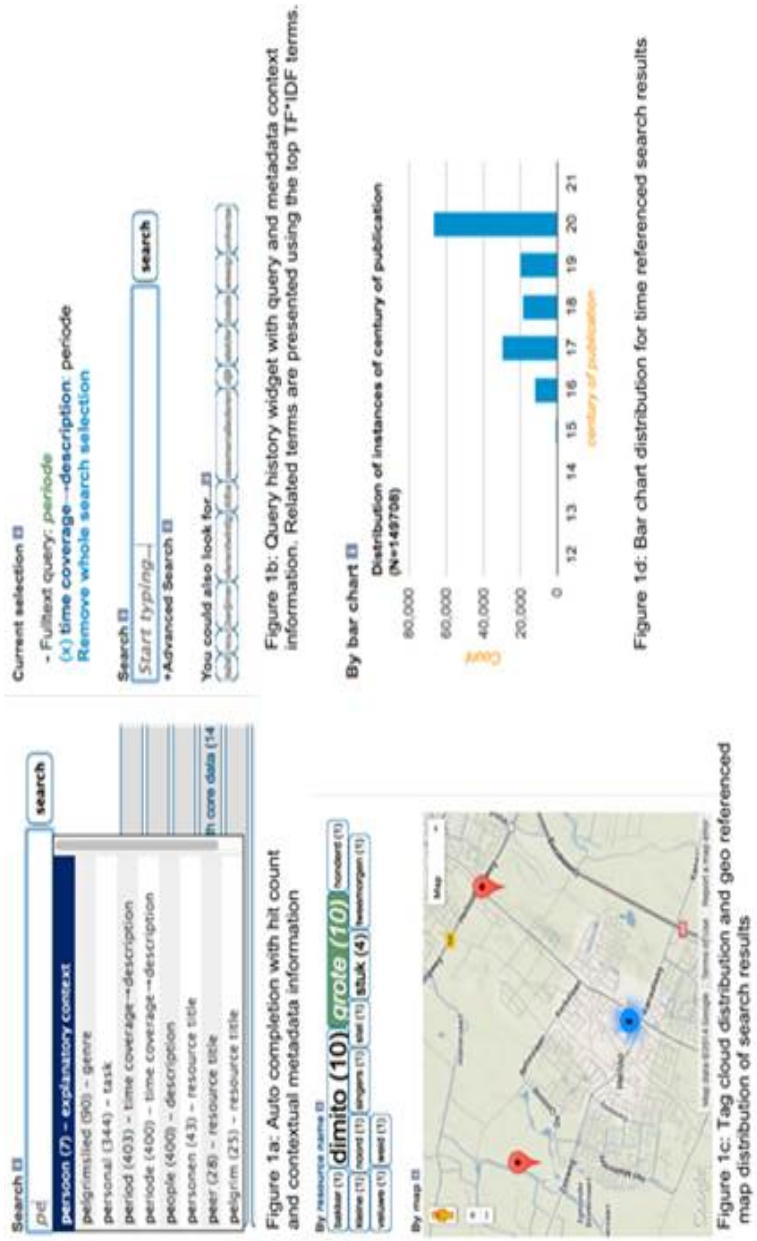

Figure 1: Meertens CLARIN Metadata Search Interface

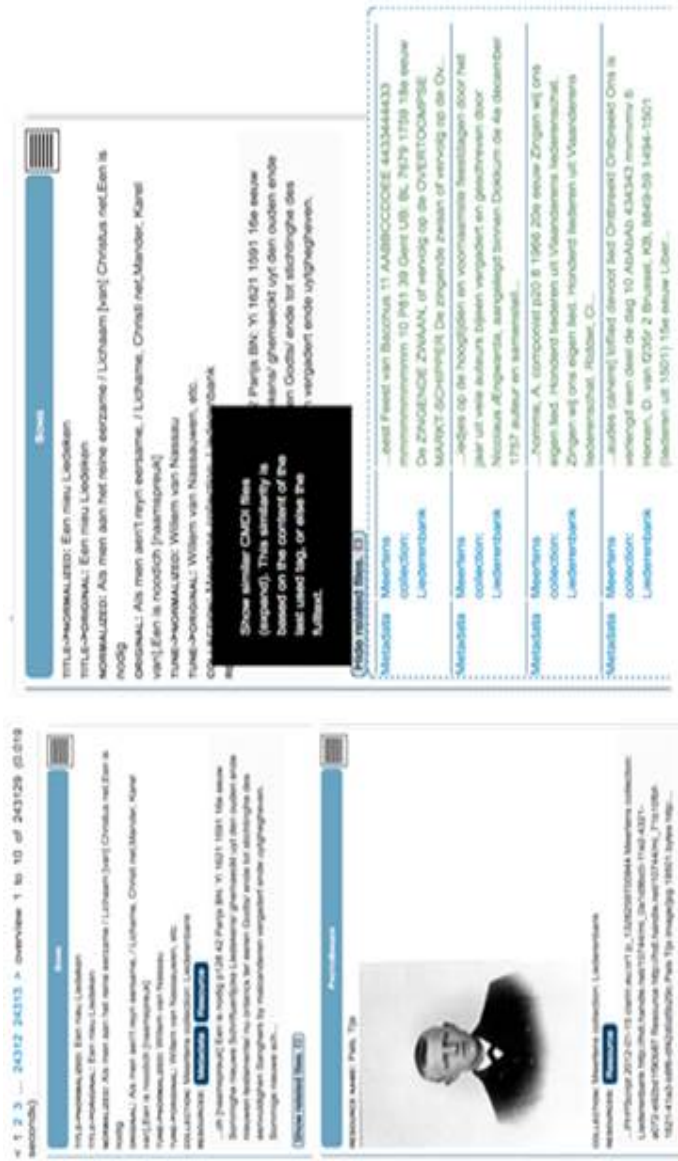

Figure 2: Meertens CLARIN Metadata Search Interface: recommended resources

system: DANS uses its own EASY system, which is built on Fedora, INL uses DSpace, MPI / TLA uses LAT, and Huygens ING and Meertens Institute use their own systems.

Some centres use special software so that users can store resources in the repository, e.g. LAMUS is used by MPI/The Language Archive, and EASY by DANS. Storing resources in the repository must be supported by software tools, since it is not an easy matter. Typically PIDs are assigned in this stage, usually to a large set of resources: a PID must be generated for the resource, it must be associated with the resource location, it must be added to the resource description, which now can be finalized and get its own PID. Provisions for legal or ethical restrictions must be taken care of. Finally the resource itself must be stored on a server that is accessible from outside of the CLARIN centre, and its resource description must be put on a location where it can be harvested by the OAI-PMH protocol.

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 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, for such services the data have to be transformed into different formats, enriched with indexes, etc. to make the search service 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. Each CLARIN centre has to have a procedure for dealing with such cases.

Each CLARIN centre must ensure long term preservation of your resource: it must not only store it on one of its servers but it must make sure that it is still accessible after e.g. ten or twenty years. 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 managing its data and for ensuring long term preservation, and work according to this procedure. These are ingredients (guidelines 6,7, and 8) of the Data Seal of Approval (DSA), which each CLARIN centre must be awarded if it is to become a certified CLARIN Centre.[32]

The DSA guidelines are elaborations of a small number of criteria that data must meet: the data can be found on the Internet; the data are accessible (clear rights and licences); the data are in a usable format; the data are reliable; and the data are identified in a unique and persistent way so that they can be referred to.

All CLARIN centres in the Netherlands have been awarded the Data Seal of Approval[33] and most are CLARIN-certified centres.[34]

[32]This DSA consists of 16 guidelines for the curation of data, 3 of which apply to the data producer, and 3 to the data consumer. The remaining 10 guidelines apply to the centre.

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

[34]See `http://www.clarin.eu/content/certified-centres`. DANS has postponed its CLARIN-certification because it needs more time to incorporate the CLARIN-specific requirements in its overall procedures, which cover a larger community than the CLARIN community. It aims to meet all requirements

5.1 Login

If a user wants to get access to CLARIN data or services, CLARIN must, for certain data and services, identify who the user is (*authentication*) and determine what the user is allowed to do (*authorization*). Systems that take care of this are therefore called *Authentication and Authorization Infrastructures (AAI)*. Both aspects will be discussed in separate sections.

5.1.1 Authentication

Authentication, i.e determining who a user is, is usually done by requiring login. Hiding resource behind a login in the CLARIN context is intended 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 that have been uploaded by a user to apply a service to it, as well as the data that result from this service. In such a case we want to make sure that only the relevant user (or a small group of users, e.g. from the same research team) see and can manipulate these data, and the user does 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?

The AAI-system mostly used in CLARIN is Shibboleth. It is a system for so-called *Federated Login* and it avoids the problems mentioned above. The basic idea is that a user, when (s)he logs in, is redirected to his/her own organisation, logs in there with the user name and password of his/her organisation, and when this is successful, the organisation communicates to CLARIN that this is a trusted user (without passing on any confidential information) who can be given access to CLARIN data and services.

I describe here globally how this works and what had to be done for it to make it work.[35] All the work done here has been carried out in the context of the IIP project and by Huygens in its eLaborate project, unless indicated otherwise.

When a user logs in on a CLARIN service, (s)he must be directed to a login with his/her own institute. For this to work, a number of things are required.

First a *CLARIN Service Provider Federation (SPF)* must be set up: this is the federation of centres that offers CLARIN services. This was done by CLARIN-PP.[36]

for certification by 2015.

[35] See Odijk (2014) for a more detailed illustration purely from the perspective of using this functionality.

[36] The CLARIN Service Provider Start-Up Federation agreement was initially concluded by a limited number of organisations, and new service providers joined later. This agreement already contains models for agreements between the CLARIN SPF and national network organisations (SURFNet in the Netherlands, DFN-AAI in Germany, etc.)

Second, an agreement must be made between this CLARIN SPF and SURFNet, so that the CLARIN SPF is recognized by SURFNet, and a trust relation is created between these parties.[37]

Third, the centre where the data reside or the service runs must be a member of the CLARIN SPF, and thus be bound by the agreement between the CLARIN SPF and SURFNet. This is necessary, because the user must be sure that (s)he is indeed using a CLARIN service, and not some unknown service that might make abuse of the situation or implicitly charge costs to the user or his/her institute. All CLARIN centres in the Netherlands are members of the CLARIN SPF.

Fourth, the organisation of which the user is an employee or student must enable the usage by its employees/students of services offered by members of the CLARIN SPF. The Netherlands has a so-called opt-in system: no service can be used by an organisation unless explicit permission has been given for it by this organisation.[38]

Requiring explicit permission for each service offered by CLARIN is not feasible, and not scalable. Fortunately, it was agreed with SURFNet that an organisation could give a single permission for the use of the whole set of CLARIN services.

SURFNet made an explicit request to all participating research organisations in the Netherlands to give their employees and students permission to use the CLARIN services. Technically, this requires a few small adaptations in some registries: it is not difficult and involves little effort. However, the response on this request was minimal: only a few organisations gave permission and made the required technical adaptations.

In order to improve the situation, a new strategy was followed. First, a package was prepared to explain exactly what was involved and what was technically required. It included a letter by the general director of NWO, in which each organisation was requested to give permission for usage of the CLARIN services. It referred to the original letter sent by SURFNet. It also included a link to a tool to test, after permission was given, whether the technical implementation of this permission actually works correctly. Second, for each organisation, a prominent researcher (mostly a full professor) active in CLARIN was approached for assistance. The idea was that a request from a prominent researcher from inside the organisation might have more success than a general request from SURFNet, and might easier lead to follow-up, face-to-face contacts, etc. The package contained a model letter that the prominent researcher had to adapt slightly to his/her own organisation, but that describes exactly what the request is, and what it involves. This approach was indeed more successful, though in some cases it still took quite some time before the permission was given and the technical measures were taken and tested. Currently, most CLARIN-NL partners have given permission to their employees and students to use the CLARIN services.[39]

Fifth, the CLARIN centre where the data reside or the server runs must implement a

[37] And similarly, in other countries for the local SURFNet counterparts there. Otherwise, Dutch researchers cannot get access to services outside of the Netherlands, and foreign researchers cannot get access to services in the Netherlands.

[38] The alternative is opt-out: each service can be used by default unless an organisation explicitly excludes its use.

[39] The same package has been sent to our colleagues in Flanders, and apparently it has been successfully used there as well, since at least some universities in Flanders (e.g. Ghent University) have access to the CLARIN services through the CLARIN AAI system. This was especially important for INL, which, by its very mission, serves both the Netherlands and Flanders, and has most of its corpus and lexicon search engines behind a login.

running version of Shibboleth (or similar software), and ensure that access to the data or service always leads to the shibboleth system, so that the credentials of the user can be checked. SURFNet offers services in this respect through SURFCONEXT. However, this service is by default accessible for researchers from the Netherlands only, which is too limited in the CLARIN context, which aims to provide access to all European researchers (and even wider). Some centres ran into this problem (Huygens ING, DANS) and have to solve it.

Sixth, the system must determine somehow, when a user logs in, to which organisation the user has to be redirected. The system does not know this, and therefore has to ask the user. A simple way to achieve this is to present the user with a list of all organisations, so that the user can make the selection.[40] But since hundreds of organisations will be in that list, doing only this is not really user-friendly.[41] Therefore additional systems are used. In particular, systems are used that put the organisations that are geographically close to the user at the top of the list: it does this by determining the user's geographical location, e.g. using HTML5 Geo Location. If you are working at your institute, it will most probably end up in the top of the list[42], and you do not have to search through the whole list with hundreds of entries. Of course, this will not work when you work at a different location. But these systems also make it possible to remember choices you made earlier, e.g. via cookies on your computer. So your institute will be in the top of the list even if you work from a different location, provided you work on the same computer.

When the user is redirected to his/her own institute, (s)he can login with his/her institute's user name and password. If the login is successful, the institute server confirms that the user is a trusted person, and (s)he can enter this part of the CLARIN infrastructure. The user's institute does *not* pass on any sensitive information such as your password.

If a logged-in user now goes to another part of the CLARIN infrastructure that requires login, it 'knows' that you are already logged in and a trusted user, so you do not have to do this again: therefore this is called *Single Sign On (SSO)*

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 when 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.

5.1.2 Authorisation

Authorisation means determining what a logged in user is allowed to do. For example, in some applications some users can only view certain data, while others are also allowed to edit them, and again others are also allowed to delete them. CLARIN centres must make provisions for such cases.

[40] Most centres use Discojuice for this purpose, but some use other systems.

[41] Having the user type in the name of the organisation is also not user friendly, and it will not be easy to make it work since usually many different variants of an organisation's name are in use, and very few people know the official name of their organisation (and the official names tend to be long, so are difficult to type without errors).

[42] Not necessarily at the top, since the accuracy of geo location systems may differ depending on the equipment you work on.

In services and applications, authorisation is usually dealt with at the service or application level, and there is no role for CLARIN. For data, CLARIN should offer provisions, but for aspects as viewing, downloading, editing and deleting so far all users have the same rights: all resource descriptions can be viewed and downloaded by all users, most resources can be viewed (often through a specific application) by all users, and some resources can be downloaded by all users. Editing and deleting is only allowed for managers of the data at the CLARIN-centre.

A special case concerns legal and ethical restrictions. Each CLARIN centre must make provisions for this, 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 and always will be 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; (3) signing a special license agreement may be required.

Some CLARIN centres in the Netherlands have special provisions to deal with such matters, e.g. the TLA/MPI. For example, one option is to show a user a text with usage conditions (but the user can access the data without reading this text). A second option shows such text but requires confirmation by the user that the text has been read and agrees to it; A third option is to require explicit permission from the data provider for usage of the data according to a specific license agreement (this is the case for, for example the IPROSLA data set, which requires special provisions to protect the privacy of the participants, who come from the (small) sign language using community in the Netherlands). Most other have arranged such matters either by providing access to such data in limited ways. For example, most text corpora at INL can only be accessed via specific search interfaces, and after login. Export of the results of the search queries is highly limited. Downloading these text corpora is simply not possible.

6 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, portals have been set up for CLARIN. The idea is that from these portals all functionality offered by CLARIN can be accessed.[43]

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 4.5.1), 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. The CLARIN portal should also provide links to national CLARIN portals.

[43]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 one portal, the CLARIN ERIC portal, should contain links to all other portals.

In addition to the Europe-wide portal, also national CLARIN portals are being created. 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 (by the CLAPOP project) and can temporarily be accessed via the URL `dev.clarin.nl` and later via `www.clarin.nl`.[44]

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 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 3.

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

7 Apply software to data

There is a lot of software in the CLARIN infrastructure that can be applied to data. Most of these have been developed in the context of resource curation and demonstrator projects and will not be discussed here. The relevant applications 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 applications will be discussed here: searching in data through *federated content search* (section 7.1), the *TTNWW* work flow system for web services (section 7.2), and *annotate-and-search* applications (section 7.3).

7.1 Federated Content Search

Most CLARIN centres maintain dedicated search engines at the level of individual resources. However, these search interfaces are often not directly accessible through web service interfaces and display a large variety of query languages and implementation details. For research infrastructures such as CLARIN making these unrelated and partly overlapping content search engines available to the research community a general perspective of these content search engines must be developed.

Federated Content Search (FCS) is a technique that may serve this purpose: FCS enables a user to enter a single query, which is sent to multiple search engines at different CLARIN Centres, each of which enables search in a specific resource with its own idiosyncratic structure and format.

FCS basically works as follows: The user wants to make a query. Of course, such a query must be formulated in some language. Federated Search in CLARIN uses the Contextual

[44] 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`.

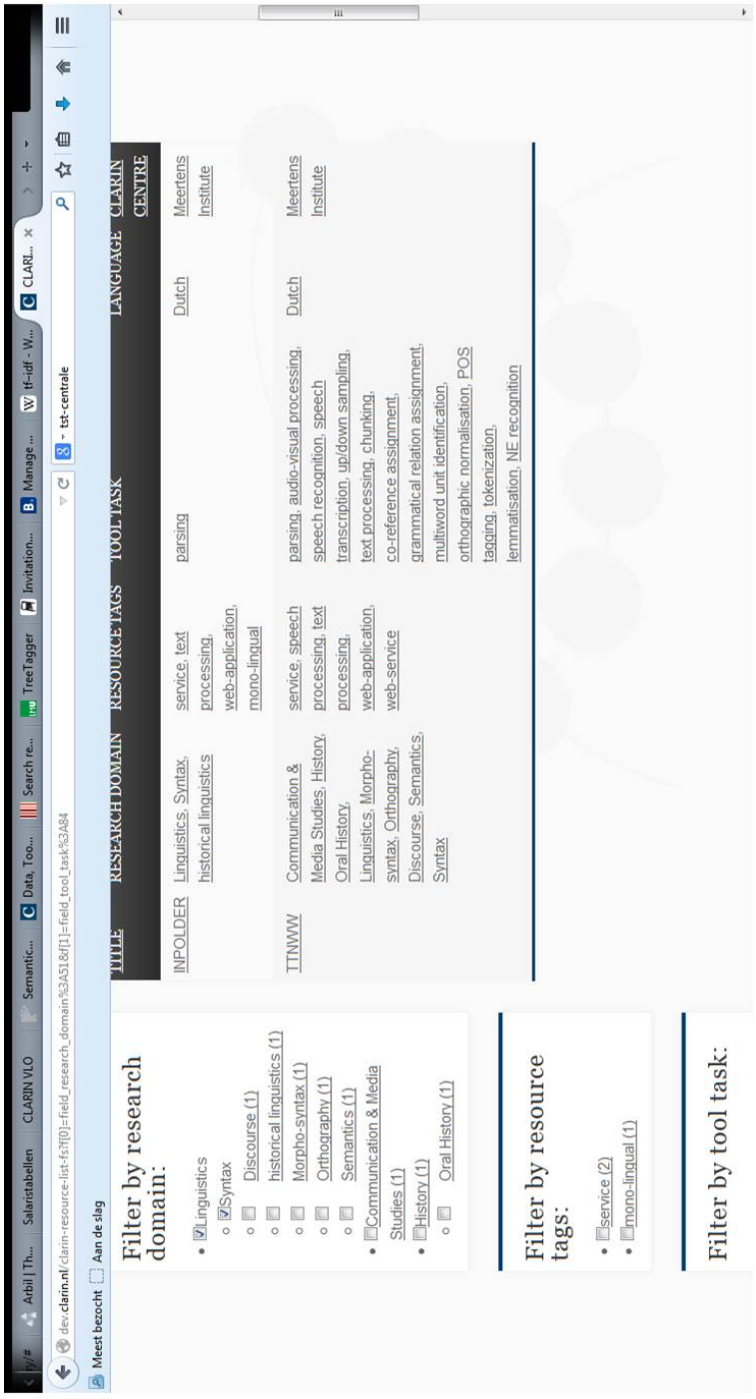

Figure 3: Selection of serviced via faceted browsing in the Dutch portal

Query Language (CQL)[45] for this purpose.

This query has to be sent to each search engine at the CLARIN Centres via some protocol. The protocol used is based on the Search Retrieval via URL (SRU) protocol, which, however, was extended by the S&D project in collaboration with the European CLARIN developer community. Extensions were needed because the local search engines cannot handle these queries and this protocol by themselves.[46] For this reason, a so-called *end point* was created for each search engine which can receive queries via the SRU protocol and translate a CQL query into a query suited for the search engine. The results of the query are of course in the format provided by the search engine. These results must therefore be translated by the end points to a common result format. Such a result format has been defined, and it is extensible. With the results from the different search engines all in a common format, they can be put together (aggregated) and presented to the user.

In order to test the approach, each CLARIN centre involved in the S&D project had to set up some end points, and they did: DANS for the Lieffering CQL Searchable database (Eighteenth-Century Music and theatre advertisements from the 's-Gravenhaagsche Courant and Gazette de La Haye), INL for the Corpus Gysseling, and for the Brieven als Buit corpus (17th and 18th century Dutch letters), MPI for the TROVA Search engine[47], Meertens for MIMORE (Morphosyntactic variations in Dutch dialects)[48]

The *S&D* project aimed to provide a combined metadata/content search solution to the end users. Through this, end users can search a central metadata catalogue and at the same time have the possibility to search through the content located at the participating CLARIN centres. To provide a single point of access to end users the CMDI metadata search engine described in section 4.5.2 was to be combined with the content search end points. For this purpose, the content search end points may be added to the metadata specification of individual resources to indicate the availability of a content search end point for this resource.

When such a content search engine is available for a specific resource, it is made accessible through the Meertens resource description search engine, which is able to detect the availability of such a service and integrates an additional widget to the user interface allowing end users to search the underlying resources directly.[49]

Both CLARIN-NL and the German national CLARIN project *D-SPIN* have adopted the CLARIN SRU/CQL protocol as a joint specification for content search end point implementations. However, some noteworthy differences between the approaches of the two projects exist. While development of a resource description search engine was part of the Dutch CLARIN-NL project, the German D-Spin project chose only to implement individual content search end points. No effort was made to provide an integrated search solution

[45] Not to be confused with the Corpus Query Language, which has the same acronym and is also highly relevant in the CLARIN context.

[46] See the discussion paper Federated Content Search for a description of the approach to federated search in CLARIN.

[47] TROVA itself searches at MPI / TLA in multiple resources, which may be in a wide range of different formats.

[48] which itself provides access to 3 databases of Dutch dialects: the Dynamic Syntactic Atlas of the Dutch Dialects, the Diversity in Dutch DP Design database, and the Goeman, Taeldeman, van Reenen Project database

[49] The VLO also allows to search for data connected to federated content search only. Currently, selecting this option reduces the number of metadata records to around 63 k records (or about 10% of the total)

for resources and their resource descriptions in this project. Instead, an aggregator called CLARIN-D Federated Content Search was developed for distributing the content search query over a number of content search engines and displaying the results. Content search end points have to be registered as part of the centers registry thus connecting content search engines to organizations rather than resources. Since organizations usually maintain specialized content search engines for various resources this makes it difficukt if not impossible to focus the content search on only a limited number of resources of interest. The CLARIN-NL approach uses the resource description to specify the content search engine end point instead, and thus establishes a relation between the content search engine and a resource. Since organizations usually maintain multiple content search end points for different resources it is therefore possible to target only specific search end points depending on the results returned by the resource description search engine. By registering content search end points in the centers registry this option is lost. It is also technically possible to combine all end points in an aggregator, as was done in the D-Spin project, by extracting all unique end points specifications from the CMDI documents. Although an aggregator was realized as a proof of concept, implementation during the project this was not pursued any further awaiting convergence at the European level concerning the registration of end points (centers registry versus resource descriptions). For the moment, CLARIN-NL has decided to follow the German approach, though as far as I can see there is no reason why the Dutch approach cannot be taken as well.

A limited form of federated content search is currently possible in data from CLARIN-NL, CLARIN-D and the Czech CLARIN project LINDAT via the CLARIN-D Federated Content Search web application. 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. It returns search hits in the form of a *KeyWord In Context (KWIC)* list.

The CMDI search engine technology developed in the CLARIN-NL S&D project has found practical application in a number of subsequent projects, including the Nederlab project and Dutch Songs Online. Although these projects take an aggregated content search approach (i.e. content is stored centrally as part of the index) rather than a federated content search approach the technological foundation in these projects is largely the same. The results of the S&D project thus demonstrate a practical applicability beyond the CLARIN domain and continue to be developed for more advanced use cases.

7.2 TTNWW Work Flow System for Web Service

The Dutch-Flemish STEVIN programme and earlier projects and programmes have resulted in a wide variety of natural language technology tools for enriching Dutch language resources with all kinds of annotations. It includes tools for orthographic normalisation, tokenizing, lemmatizing, part-of-speech tagging and parsing, for assigning semantic roles, for assigning co-reference relations, and for identifying and analyzing named entities, as well as tools for the automatic orthographic transcriptions of the speech in audio files.

Unfortunately, these tools can only be used by experts. In addition, the relevant tools were mostly developed separately, so that each uses its own input and output formats, hampering interoperability.

In order to address this problem, the Netherlands and Flanders started up a cooperation

project (called *TTNWW*) with the aim of making it possible for arbitrary humanities researchers to use these tools. To that end, each of these tools was turned into a web service, and usually combined with other web services in a web services work flow, so that a user just has to upload the relevant resources and click on a button to have the resource enriched with annotations. An explicit additional goal was to develop a common format for textual resources to further improve ease of use and interoperability.

In order to make such work flows, a work flow system was used: the TTNWW project opted for TAVERNA. TAVERNA enables users to make their own work flows, but the TTNWW project has yielded only a number of fixed work flows, since it was expected that creating one's own work flows was not among the requirements of humanities researchers and was too difficult for most of them.

Web services are software programmes that communicate with other services over the web. Communication between programmes is usually formally defined in an Application Programmers Interface (API). 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. Most tools for TTNWW were converted into web services by a piece of software called Computational Linguistics Application Mediator (CLAM), which turns desktop software into a web service using the REST protocol.[50].

One of the results of the TTNWW project was also a new format for textual resources called Format for Linguistic Annotation (FoLiA). It defines a format, comes with two libraries, and a web application for annotating FoLiA resources called FoliA Linguistic Annotation Tool (FLAT) is under development.

Though FoLiA has been very successful (many other resources have adopted this format), it is still not used in all TTNWW work flows, whether for input or for output, and its design clashes with the design of the LASSY treebank formats. As a consequence, though steps in the right direction have been made, there is still a lot to be done here to improve interoperability. A few small projects (e.g. *OpenConvert*, INL) are being started up in 2014 to address some (but certainly not all) of these issues.

In addition, though FoLiA may become an important standard format for Dutch textual language resources, it is not yet recognized as a supported format at the European CLARIN level,[51] and of course it differs from formats used in other countries. So, interoperability of textual resources for different languages is still a matter of concern.

7.3 Annotate and Search Applications

Creating annotations for language resources is useful, but for many users it is not enough: they want to be able to search and analyse these annotated resources. Though there are several search and analysis tools for linguistically annotated corpora in CLARIN-NL, each is tied to a specific resource or resource set. Examples are the Groningen word relations search application (applies to the LASSY treebanks), the INL search application for the Contemporary Dutch Corpus, and GrETEL (applies to the LASSY and CGN treebanks).

[50]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

[51]One of the reasons for this is that there is no procedure for this in CLARIN, something that, I hope, will change soon.

It is currently not possible for a user to use these search and analysis applications for his/her own text corpus that has been enriched with annotations by e.g. TTNWW. A number of projects are currently being set up to create this functionality, among them *Parse and Query* (PaQu, Groningen), and *Autosearch* (INL), and they are expected to provide this functionality early 2015.

8 Concluding Remarks

@@write

Acknowledgments

The sections 4.5.2 and 7.1 are slightly rewritten versions of sections from the S&D final report written by Marc Kemps-Snijders. 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).

[Broeder, 2009] Daan Broeder. LAMUS/LAT repository system. Presentation, Max Planck Institute Nijmegen, 2009.

[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.

[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, 2009] Jan Odijk. Data categories and ISOCAT: some remarks from a simple linguist. Presentation given at FLaReNet/CLARIN Standards Workshop, Helsinki, 30 September 2009.

[Odijk, 2014] Jan Odijk. The CLARIN infrastructure in the Netherlands: What is it and how can you use it? unpublished article, Utrecht University, August 2014.

[Van Eynde, 2004] Frank Van Eynde. Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands. CGN report, Centrum voor Computerlinguïstiek, KU Leuven, Leuven, Belgium, February 2004.

[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.

[Ďurčo and Windhouwer, 2014] Matej Ďurčo and Menzo Windhouwer. The CMD cloud. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 687–690, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[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.