PART I

Technical Infrastructure

CHAPTER 3

Introduction to the CLARIN Technical Infrastructure

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

ABSTRACT

This chapter provides an introduction to the design of the CLARIN technical infrastructure, with a focus on the Netherlands part. It provides a basic introduction to the techniques behind PIDs, CMDI-metadata, authentication and authorisation (AAI), semantic interoperability related to CMDI-metadata, and search. Search covers searching for data through metadata with the VLO and the Meertens metadata search application, as well as federated content search activities in the Netherlands. The chapter ends with an introduction to the chapters of Part I on the technical infrastructure of CLARIN.

3.1 Introduction

This chapter serves as an introduction to Part I of this book, which covers CLARIN's technical infrastructure, and it also provides an introduction to the design of this technical infrastructure, with a focus on the Netherlands part. I will try to explain what has to be done behind the scenes to make the CLARIN infrastructure in the Netherlands work. For many aspects, a more detailed description is required than can be given in this chapter. For these, I refer to other chapters in Part I of this book.

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

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

How to cite this book chapter:
Odijk, J. 2017. Introduction to the CLARIN Technical Infrastructure. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 33–44. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.3.

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 analyse research data (*demonstrators*). The focus of the current part of this book (Part I) 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. The two most important and largest subprojects were:

IIP (*Infrastructure Implementation Project*), which has implemented the basic functionality of the CLARIN infrastructure in the Dutch CLARIN centres. It will be described in more detail in chapter 4.

S&D (*Search & Develop*), which has developed a metadata search application and worked on federated content search. Its results will be described in section 3.6.2 and section 3.6.3.

They were complemented by a whole range of smaller subprojects that we will not all mention here. See `http://www.clarin.nl/node/281` for more details on these projects. Some of these projects will be mentioned in the course of this chapter and some have their own chapter in this part of the book.

The construction of the Netherlands part of the CLARIN infrastructure requires a whole range of activities. I mention them here and indicate where they will be discussed. Some are discussed in this chapter, others in other chapters of this part of the book (Part I).

- Setting up a network of certified CLARIN centres and contributions to a variety of Type A services and registries, e.g. applications and registries for supporting the creation of CMDI metadata (see chapter 4).
- Setting up a Persistent Identifier (PID) infrastructure (see section 3.2 and chapter 4).
- Providing and testing metadata profiles and components (see section 3.3 and chapter 4).
- Setting up a CLARIN-compatible Authentication and Authorisation Infrastructure (AAI) (see section 3.4 and chapter 4).
- Contributions to further development and maintenance of facilities for semantic interoperability (see section 3.5 and chapter 4).
- Setting up a metadata search and browse application (see section 3.6.2).
- Setting up a Federated Content search application (see section 3.6.3).
- Setting up a CLARIN-NL Portal (see section 3.7).

In addition, the Low Countries collaborated in setting up a system for organising language technology web services in a workflow system (TTNWW), which will be discussed in chapter 7.

3.2 Persistent Identifiers

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 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)* 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 (issuing), the assignment, the maintenance and the resolution of PIDs.

A persistent identifier is an identifier, ideally without any internal structure or semantics. It is is created and issued by a PID-service organisation. It is issued to the organisation that requested the PID and considers itself responsible for it.[2] A newly created PID must be unique. A PID is associated with a URL by the requesting organisation (PID assignment), 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 assigned 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.

In the CLARIN infrastructure, each metadata record[3] is assigned a PID. In this way, a user or software program that wants to use a specific resource can simply refer to the PID assigned to its associated metadata and never has to change this reference anymore. The PID resolution system will resolve the PID, i.e., in the context of accessing web resources, replace it by its associated URL, and transfer the user or the software to the metadata record, and, through this metadata record, to the resource itself.

CLARIN requires the use of the Handle System PID-technology. An example of a Handle PID is *10032/12824827a77b9602cc66840a62aedf43*. The uniqueness of each PID is guaranteed because each issuer has its own prefix (10032 in this example), and the PID-system guarantees the assignment of a unique new identifier within the system. Having a PID preceded by the prefix `http://hdl.handle.net/` turns it into a URL. By clicking on it, it brings the user to the PID, which is resolved and leads the user to the metadata record it is associated with.

Chapter 4 will describe in more detail how the various centres in the Netherlands dealt with PIDs, their assignment, and their resolution.

3.3 CMDI Metadata

Metadata play a crucial role in CLARIN in offering services for finding data and software. Metadata in CLARIN must be in CMDI-format. CLARIN-NL made many contributions to the CLARIN CMDI infrastructure.

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 and tested profiles for text corpora, lexical resources and speech corpora, and for a number of specific other resources in the Netherlands. This was done initially only for data, but not for software. Originally, very few metadata for software were made in CLARIN-NL, 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 five pieces of software curated in CLARIN-NL. It is currently being refined and applied to all software curated or created in CLARIN-NL.

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

CMDI offers, through its flexibility, many advantages, but this flexibility also has some drawbacks. Flexibility is needed when 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 exist so that reuse of profiles

[2] The issuing organisation and the requesting organisation can be identical, but this need not be the case.

[3] Usually a resource itself is assigned a PID as well, though this is not required.

and components can be maximised, and unnecessary errors or omissions can be avoided. It also provides 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.

The Component Registry was 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 one's resource, since the registry consists, in essence, of a flat list of profiles and components, and advanced 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. The lack of a clear versioning strategy also increased the problem.[4]

In 2014, a project has indeed been started up to investigate how the quality of existing and new metadata can be improved, how reuse of existing profiles and components can be increased, and how profiles relate to one another. It resulted in a report on a strategy for metadata quality (Kemps-Snijders, 2014). The problem is not unique to the Netherlands. Austria has reported it as well, and has developed a tool, the SMC-browser, to investigate the relations between profiles and components (Ďurčo and Windhouwer, 2014). The CLARIAH-CORE successor project will address these issues as well.

3.4 AAI

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 (*authorisation*). Systems that take care of this are therefore called *Authentication and Authorisation Infrastructures (AAI)*. Both aspects will be discussed in separate sections.

3.4.1 Authentication

In chapter 2 we saw that it is in some cases necessary or desirable to authenticate a user, i.e to determine who the user is. Authentication is usually done by requiring login.

Logging in in the CLARIN infrastructure is not an obvious thing, as we described in chapter 2. We repeat here the major problems: the CLARIN infrastructure is a distributed infrastructure, spread out over all of Europe (and beyond), so how can it be avoided that the user has to login again each time a resource happens to be located at a different centre? How can it be avoided that the user has 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 for a user community as large as the CLARIN one? Clearly, it does not scale to have every centre use a separate identity store.

The basic idea behind the solution adopted in CLARIN is that a user, when (s)he logs in, is redirected to his/her own organisation (which acts as an identity provider), logs in there with the user name and password of the organisation, and, when this is successful, the organisation communicates to CLARIN that this is a trusted user, who can be given access to CLARIN data and services. Since every user is directed to his/her own identity provider, this type of system is called Federated Identity Management (FIM).

We describe here globally how this works and what has to be done for it to make it work. The work that had to be done here has been carried out in the context of the IIP project (unless stated otherwise) and is described in chapter 4.

[4] For example, at a certain point there were three different components called *GeneralInfo* created by user *nalida*, and it was totally unclear how they were related. Currently (November 2016) there is fortunately only one (but many components with the same name by other users).

When a user tries to log in on a CLARIN service, (s)he must be directed to a login at his/her own institute. For this to work, a number of things are required, which are partly administrative and partly technical in nature.

- First, a *Service Provider Federation (SPF)* must be set up: this is a federation of centres that offer software services. CLARIN set up its own CLARIN SPF. This was done by CLARIN-PP.
- Second, an agreement must be made between this CLARIN SPF and the National Research and Education Network (NREN), i.e. SURFnet in the Netherlands and Belnet in Flanders, so that the CLARIN SPF is recognised by the NREN, and a trust relation is created between these parties.[5]
- 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 abuse 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 a member of an organisation unless explicit permission has been given for it by this organisation.[6]
  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.
- Fifth, the CLARIN centre where the data reside or the server runs must implement a 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). Making FIM available in the European CLARIN context requires some additional configuration and administrative actions.
- 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.[7] But since hundreds of organisations will be in that list, doing only this is not really user-friendly.[8] 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 the user is working at his/her institute, it will most probably end up in the top of the list,[9] and the user does not have to search through the whole list with hundreds of entries. Of course, this will not work when the user works at a different location. But these systems also make it possible to remember choices a user made earlier, e.g. via cookies on the user's computer. So the user's

[5] And similarly in other countries with their local NRENs. Otherwise, Dutch researchers cannot get access to services outside of the Netherlands, and foreign researchers cannot get access to services in the Netherlands.

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

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

[8] 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).

[9] Not necessarily at the top, since the accuracy of geolocation systems may differ depending on the equipment one works on.

institute will be in the top of the list even if the user works from a different location, provided the user works 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.

If a logged-in user now goes to another part of the CLARIN infrastructure that requires login, this other part 'knows' that this user is already logged in and a trusted user, so (the user does not have to do this again. In this way, *Single Sign On (SSO)* is implemented.

All this requires a lot of communication between various systems. AAI in CLARIN uses SAML[10] Version 2.0 for this and it is therefore called SAML-based (or SAML2-based) FIM.

3.4.2 Authorisation

Authorisation means determining what a logged in user is allowed to do. For example, in some applications some users are only permitted to view certain data, while others are also allowed to edit them, and again others are also allowed to delete them. Though CLARIN imposes no requirements here, CLARIN centres must of course ensure proper use of their resources, so they must make provisions for such cases.

In services and applications, authorisation is usually dealt with at the service or application level, and there is no role for CLARIN. For data, all users have the same rights for aspects such as viewing, downloading, editing and deleting: all metadata 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 get access to resources that have such restrictions actually get access to them, and to ensure that researchers use resources in the way they are allowed to use them. Restrictions related to the Creative Commons licensing conditions must be explicitly marked in the metadata and are visualised in the VLO with the Creative Commons 'laundry tags' to inform the users of their rights and obligations. 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; or (3) signing a special licence agreement may be required.

Some CLARIN centres in the Netherlands have special provisions to deal with such matters, e.g. the MPI/TLA. For example, one option is to show a user a text with usage conditions, but let the user 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 agreed to. A third option is to require explicit permission from the data provider for usage of the data according to a specific licence agreement (this is the case for, for example, the IPROSLA dataset, which requires special provisions to protect the privacy of the participants, who come from the (small) sign language using community in the Netherlands). Other centres have arranged such matters 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.

[10] Security Assertion Markup Language.

3.5 Semantic Interoperability

The flexibility of CMDI is only possible if the semantics of the metadata elements is made explicit. Explicit semantics for a resource or metadata is obtained by explicitly linking each element and its possible values in the resource and the metadata to an element of a CLARIN-recognised concept or data category registry. The most prominent registry for semantic interoperability in CLARIN until 2014 was 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 a 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 programs.

ISOcat is basically just a flat list of data categories.[11] Finding an existing data category that might be relevant is therefore 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)? One can only find these by manually going through the whole list of data categories. In part 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*[12]) as identical or near-identical. This can be done in a special registry, called RELcat (which, however, never got beyond α-version status).

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 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 unanalysable values. But we want to associate parts of these tags to ISOcat data categories, 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 data categories can be assigned to parts of the tag.

Finally, ISOcat may be the primary registry for semantic interoperability in CLARIN, but it is 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 SIL for ISO). In order to use such other registries in

[11] There is a little bit of hierarchy in it through so-called complex categories which basically group a limited set of simple types, and there is a division of the data categories by thematic domain, but this is by far not enough for efficiently finding closely related data categories.

[12] And the category with name *syntactic function*.

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. The CLARIN Vocabulary Service is described in more detail in chapter 5

In 2013, CLARIN switched from ISOcat to the CLARIN Concept Registry (CCR; see chapter 4 for the background. CCR is a concept registry according to the W3C SKOS recommendation (Schuurman et al., 2016) and is hosted by the Meertens Institute. It has not really played a role in CLARIN-LC, but it will be important in CLARIAH-CORE.

3.6 Search

If there is a CMDI metadata record for each resource, and if the metadata can be referred to via a PID (and the resource itself via the metadata), combined with harvesting facilities through OAI-PMH, everything is in place to create functionality for browsing and searching for resources.

This functionality requires a browsing and/or search engine in combination with a web interface. Such engines operate on a database of CMDI metadata located on a server of a CLARIN centre that offers essential infrastructure services (a so-called Type A centre), which is filled and regularly updated by metadata harvesting, as will be described below.

CLARIN offers the *Virtual Language Observatory (VLO)* as a browser and search application to search for resources via their metadata. It will be discussed in section 3.6.1. CLARIN-NL developed the *Meertens CLARIN Metadata Search Application*, which will be discussed in section 3.6.2.

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 3.7.

3.6.1 VLO

As described in chapter 2, the Virtual Language Observatory (VLO) offers facilities for browsing and searching in CMDI metadata. It enables a user to do a string search for keywords that occur in the metadata, and it offers faceted browsing.

In order to make this functionality possible, the Type A centre that hosts the VLO regularly gathers the metadata of all CLARIN centres in one central database. This process is called *metadata harvesting*, and it is done through the OAI-PMH protocol. This has to be done regularly,[13] since new metadata will regularly appear at each CLARIN centre.[14]

The 'harvesting' software run by a Type A centre must 'know' where the metadata of each centre can be found. This is one of the reasons why the CLARIN centre registry has been set up.[15] 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 metadata 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 is an overview of the OAI-PMH end points.

---

[13] See `http://www.clarin.eu/faq/when-metadata-vlo-harvested` for the harvesting update schedule for the *Virtual Language Observatory* (see section 3.6.1).

[14] Currently, only the MPI / TLA does such regular harvesting. The Meertens Institute only occasionally took a snapshot of the metadata harvested by MPI / TLA for its CLARIN search engine.

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

3.6.2 Meertens CLARIN Metadata Search

The functionality of the Meertens CLARIN Metadata Search application was described in chapter 2, section 2.2.

This metadata search application has been created taking into account the diversity of the CMDI metadata descriptions and descriptive metadata elements. Harmonisation of the metadata fields using ISOcat concepts has proven to be possible and an automated ingest procedure for CMDI metadata files has been realised. The set up has been tested against all available CMDI profiles from the CLARIN EU community.

The metadata search application still exists, but it is not maintained or operated. No new metadata are harvested, so the current application contains only metadata from several years back. Development and operation of this application was stopped because it was intended as the initial step for federated content search, but development of the Dutch federated content search was stopped as well, as described in section 3.6.3.

The metadata search application has also been used as part of the Nederlab project, though a new design of the interface tuned to the intended Nederlab users was created.

3.6.3 Federated Content Search

Most CLARIN centres maintain dedicated search applications 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 a research infrastructure such as CLARIN that aims to offer an integrated search facility to make 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 Query Language (CQL)[16] 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 was originally developed in the library world for federated metadata search and was extended by CLARIN to cover textual content search.[17] 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

[16] Not to be confused with the Corpus Query Processing Language, which is sometimes also abbreviated as CQL and is also highly relevant in the CLARIN context.

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

Dutch letters), MPI for the TROVA Search engine,[18] and Meertens for MIMORE (Morphosyntactic variations in Dutch dialects).[19]

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 3.6.2 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 should be made accessible through the Meertens metadata 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.[20]

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 metadata 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 for resources and their metadata 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 centres' registry thus connecting content search engines to organisations rather than resources. Since organisations usually maintain specialised content search engines for various resources this makes it difficult if not impossible to focus the content search on only a limited number of resources of interest. The CLARIN-NL approach uses the metadata to specify the content search engine end point instead, and thus establishes a relation between the content search engine and a resource. This should be more efficient, since organisations usually maintain multiple content search end points for different resources. By registering content search end points in the centres' 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 realised as a proof of concept, implementation during the project was not pursued any further awaiting convergence at the European level concerning the registration of end points (centres' registry versus metadata). For the moment, CLARIN-NL has decided to follow the German approach, since there appears to be a critical mass of adoption of this approach within CLARIN. Although, as far as we can see, there is no reason why the Dutch approach cannot be taken as well, activities in this area stopped and the attention of the relevant researchers was shifted to Nederlab.

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. It returns search hits in the form of a *KeyWord In Context (KWIC)* list. At the time of writing one could search in resources at least from CLARIN-NL, CLARIN-D, LINDAT (CLARIN-CZ) and CLARIN Poland. See `https://centres.clarin.eu/fcs` for a full overview of the current end points. Work on federated content search is continuing, and

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

[19] Which itself provides access to three 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

[20] The VLO also allows to search for data connected to federated content search only. Selecting this option reduces the number of metadata records (at the time of measuring (2014) to around 63,000 records, or about 10% of the total).

work is underway to formulate queries in the Corpus Query Processing (CQP) language (Evert and The OCWB Development Team, 2010), both at the European level and at the Dutch national level (CLARIAH-CORE project).

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.

3.7 Portal

The functionality of the CLARIN-NL portal was described in chapter 2, section 2.5. It has been implemented as a straightforward website using the Drupal content management system. CLARIN-compatible login has been created here as well.

For the faceted browsing and searching in data and software, small taxonomies for the facet values were created. Having the freedom of experimenting with the relevant values in a working faceted browser without being penalised for not being compatible with existing defined values sets was a big advantage. However, with these taxonomies relatively stable now, we should work on including them into the CCR, and to derive the faceted browsing and search interfaces automatically from CMDI-descriptions of the relevant data and software, since it will be inefficient to maintain the faceted browsing and search on the portal directly. Deriving the faceted browsing and search from CMDI-descriptions for data and software is indeed being worked on in the CLARIAH-CORE successor project.

3.8 Concluding Remarks

I have provided an introduction to the major technical requirements that the CLARIN infrastructure must meet. In this chapter I described requirements and CLARIN solutions for persistent identifiers, for metadata, for authentication and authorisation, and for semantic interoperability. I described the options offered for search for data through VLO and the Meertens Metadata Search engine, and for federated content search, how they were implemented, and what problems were encountered in implementing them. And finally, we described the implementation of the CLARIN-NL portal.

The remainder of Part I is structured as follows. In chapter 4, a description is given of the construction of the CLARIN infrastructure in the Low Countries (mainly the Netherlands). It deals with the set-up of the network of CLARIN centres and their certification as CLARIN centres, with the set-up of essential infrastructure services, with the contributions by the Netherlands to several central CLARIN registries and services, and with how data and software have been made available to researchers via the CLARIN centres.

Chapter 5 deals with the CLARIN vocabulary service CLAVAS, a SKOS-based knowledge system to provide integrated access to a variety of managed vocabularies (copied from other knowledge sources), so that CLARIN users can transparently use concepts and data categories from each of them.

Chapter 6 deals with the FoLiA-format: a format that emerged out of the CLARIN-NL project and a range of other projects as a *de facto* standard for annotated textual resources. It discusses the format, the design principles behind the format, and the tools that accompany the format, e.g. for validation, editing, conversion, search and visualisation of documents in FoLiA-format.

Chapter 7 describes the TTNWW workflow system for web services: through this system a whole range of natural language processing tools for the Dutch language created in earlier projects were turned into web services and combined in workflows, so that any humanities researcher can enrich data in a very user-friendly manner with automatically created annotations such as lemmas, part-of-speech codes, syntactic structures, named entity annotations, etc. TTNWW covers not only textual data but also audio data, which can be enriched with automatically generated orthographic transcriptions.

Chapter 8 describes a bridge between CMDI-metadata and Linked Open Data. CMDI is the *de facto* standard with CLARIN for metadata. However, its use outside of CLARIN is very limited. In order to make the CMDI-metadata also available to the Linked Open Data community, a bridge has been made to convert CMDI-metadata into Linked Open Data in the Resource Description Format (RDF).

Acknowledgements

This work was financed by CLARIN-NL and CLARIAH.

References

Evert, Stefan and The OCWB Development Team (2010), The IMS Open Corpus Workbench (CWB): CQP Query Language Tutorial, *Ocwb report*, IMS, Stuttgart. http://cwb.sourceforge.net/files/CQP_Tutorial/.

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.

Kemps-Snijders, Marc (2014), Metadata quality assurance for CLARIN, *Clarin report*, CLARIN-NL / Meertens Institute, Utrecht / Amsterdam. http://www.clarin.nl/sites/default/files/The%20Metadata%20Quality%20Assurance-final.pdf.

Odijk, Jan (2009), Data categories and ISOCAT: some remarks from a simple linguist, Presentation given at FLaReNet/CLARIN Standards Workshop, Helsinki. http://www.csc.fi/english/pages/neeri09/workshop/materials/odijk.pdf.

Schuurman, Ineke, Menzo Windhouwer, Oddrun Ohren, and Daniel Zeman (2016), CLARIN Concept Registry: The New Semantic Registry, *in* De Smedt, Koenraad, 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 Eynde, Frank (2004), Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands, *CGN report*, Centrum voor Computerlinguïstiek, KU Leuven, Leuven, Belgium. http://www.ccl.kuleuven.be/Papers/POSmanual_febr2004.pdf.

Ďurčo, Matej and Menzo Windhouwer (2014), The CMD cloud, *in* Calzolari, Nicoletta, 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)*, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 687–690.