CHAPTER 4

Building CLARIN Infrastructure in the Netherlands

Daan Broeder[1,a], Jan Theo Bakker[b], Marco van der Laan[c], Marc Kemps-Snijders[d], Menzo Windhouwer[e], Marjan Grootveld[f]

[a]Meertens Institute, daan.broeder@meertens.knaw.nl, [b]INL, JanTheo.Bakker@inl.nl, [c]INL, Marco.vanderLaan@inl.nl, [d]Meertens Institute, marc.kemps.snijders@meertens.knaw.nl, [e]Meertens Institute, menzo.windhouwer@meertens.knaw.nl, [f]DANS, marjan.grootveld@dans.knaw.nl

ABSTRACT

In 2011 the Dutch national CLARIN project started the Infrastructure Implementation Project (IIP) as one of the means to build a national CLARIN infrastructure based on a network of CLARIN centres and to contribute to the CLARIN EU-wide effort. The IIP resulted in the certification of four CLARIN centres in the Netherlands and the provisioning of many Dutch language resources and services.

4.1 Introduction

The CLARIN Research Infrastructure is being created by a coordinated effort of many national projects and initiatives to provide an interoperable research infrastructure for the Humanities and Social Sciences on a European scale (T. Varadi et al., 2008). CLARIN has been on the ESFRI roadmap[2] since 2008 and was granted ERIC status in February 2012. CLARIN ERIC coordinates the interoperability and consistency of the infrastructure and the different national contributions through committees and task forces, although the national projects have freedom to choose their contributions to the EU infrastructure and in arranging the integration of CLARIN infrastructure with their national research projects.

[1] Daan Broeder and Menzo Windhouwer were both employed at the Max Planck Institute for Psycholinguistics (MPI) and involved in the CLARIN-NL project during the IIP; Daan Broeder was the IIP project coordinator.

[2] See http://cordis.europa.eu/esfri/roadmap.htm (accessed 14 January 2016).

How to cite this book chapter:

Broeder, D, Theo Bakker, J, van der Laan, M, Kemps-Snijders, M, Windhouwer, M and Grootveld, M. 2017. Building CLARIN Infrastructure in the Netherlands. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 45–59. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.4. License: CC-BY 4.0

Two major projects were defined in CLARIN-NL to provide basic infrastructure services: the Infrastructure Implementation Project (IIP) and the Search and Development (S&D) project. The results of the S&D project were described in chapter 3 (see also Zhang et al., 2012); the IIP project will be described in this chapter.

The IIP was intended to create the Dutch part of the CLARIN technical infrastructure, and also provided resources to the CLARIN centres to adapt their internal workflow and infrastructure to CLARIN requirements.

Adoption of this technical infrastructure by the candidate CLARIN centres must culminate in:

- their certification as a CLARIN B centre[3]
- making their resources and services available in a CLARIN-compatible way

This chapter describes how this adoption was achieved. It is structured as follows: in section 4.2 we describe the background for the CLARIN infrastructure construction. In section 4.3 we discuss the set-up of the network of CLARIN centres in the Netherlands and their certification as CLARIN centres. In section 4.4 we describe how the essential infrastructure services were set up. Section 4.5 describes how data and software services were made available by the CLARIN centres. Section 4.6 deals with their contributions to central infrastructure registries and services. We end the chapter in section 4.7 with conclusions and recommendations.

4.2 CLARIN-NL Infrastructure Building Background

The design and prototyping of the CLARIN technical infrastructure had already begun during the CLARIN preparatory phase (2008–2011), in collaboration with a number of European partners (T. Varadi et al., 2008), and the basic technical ambitions and the infrastructure design remain largely valid to this day.

The basic CLARIN technical infrastructure can be broken down into a number of key areas:

- Metadata to find resources by using the Component Metadata Infrastructure (CMDI).
- Semantic interoperability by using central semantic registry services to provide semantic interoperability for concepts used within metadata and annotations.
- Persistent Identifiers (PIDs) to identify resources and to provide options to refer to multiple copies of a resource and handle migrations in a transparent manner.
- Orchestration of web services to provide a workflow system for Language Technology (LT).
- Authentication and Authorisation Infrastructure (AAI) framework using SAML2 Identity Management for user identification and authorisation to provide a domain where researchers have a single identity and can use Single Sign On (SSO) for all CLARIN services.
- Recommended CLARIN formats for Language Resources (LRs).

This technical infrastructure should be supported and populated by a large set of European centres that provide language data and technology.

CLARIN ERIC and the many CLARIN committees and task-forces created a framework for discussing and certifying interoperability requirements. This culminated in a system of requirements for CLARIN centre types and an accompanying certification process.

Every national CLARIN project is expected to support a number[4] of certified CLARIN centres publishing useful LR data and LT services for the community. In addition, some contributions

[3] See https://www.clarin.eu/content/centres (accessed 14 January 2016) for an explanation of CLARIN centres and the assessment procedure.

[4] Currently every national CLARIN project should at least have one type B centre.

to the development and maintenance of central CLARIN infrastructure services are expected. In the following sections, we describe how, in this context, the IIP implemented the CLARIN infrastructure for the Netherlands and the contributions of CLARIN-NL to the CLARIN EU infrastructure.

4.3 A Network of CLARIN Centres

One of the main purposes of the IIP was the preparation of first four, and later five, organisations to become certified CLARIN centres. Such centres are instrumental in the provisioning of data and services in a standardised, CLARIN-compatible way; the collaboration between centres makes such provisioning more easily transferable and thus sustainable.

CLARIN centres come in different types[5]. For the Netherlands, three types are relevant: type A, type B, and type D centres.

4.3.1 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 started with four Type-B candidate centres; at a later stage one more (Huygens ING) joined. These candidate centres differ in the kind of resources that they are interested in, usually as a function of their research interests. The following is a list the Dutch Type B CLARIN centres and characterises the resource types they are most interested in:

- The Meertens Institute (MI) holds resources relevant for the study of the function, meaning and coherence of cultural expressions; and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.
- The Institute for Dutch Lexicology (INL) provides resources and services that are relevant to the lexicological study of the Dutch language; and resources relevant for research in and development of language and speech technology.
- The Max Planck Institute for Psycholinguistics (MPI/The Language Archive) houses resources related to the study of the psychological, social and biological foundations of language; documentation of endangered languages; resources for sign languages; phonetic resources for the study of phone perception; speech error databases; and also tools for creating and annotating resources.
- Data Archiving and Networked Services (DANS) provides archiving services and access to a broad classification of research data in the fields of humanities, such as oral history, archaeology, geospatial sciences and behavioural and social sciences.
- The Huygens ING Institute (HI) holds resources related to the study of the history and literature of the Netherlands.

Most centres provide not only data but also software services.[6] DANS, however, commonly favours the provision of data only. For type B centres a certification procedure was set up in CLARIN. A centre can become a certified CLARIN centre if it meets the following requirements:

1. it must offer metadata for the centres’ resources in the format used by the Component Metadata Infrastructure (CMDI; Broeder et al., 2010)

[5] See https://www.clarin.eu/content/centres, accessed Jan 14, 2016 for an explanation of CLARIN centers and the center assessment procedure.

[6] Such centres are sometimes called B+ centres, but this is not an official term

2. it must issue Persistent Identifiers (PIDs) for metadata and resources
3. it must have implemented SAML2-based Federated Identity Management (FIM) for user identification and authentication when accessing protected resources
4. it must offer resources in CLARIN-recommended formats
5. it must comply with the Data Seal of Approval (DSA)[7]

Optionally a centre can also make available one or more endpoints that are compatible with the approach adopted for Federated Content Search in CLARIN.

We will discuss the various requirements for certification one by one:

4.3.1.1 CMDI (Metadata) Resource Descriptions

Each CLARIN centre publishes many CMDI-metadata records, predominantly metadata for the centre's own resources, but in part metadata from external researchers who deposit their resources at the centre, or resource descriptions from the CLARIN-NL Data Curation Service (DCS; see chapter 2, section 2.4.5).

The resource descriptions must be made public, otherwise nobody will know of the resources' existence: each CLARIN centre makes its resource descriptions available through a publicly accessible service using the OAI-PMH protocol (*Open Archives Initiative - Protocol for Metadata Harvesting*; Lagoze et al., 2002), which allows other programmes to harvest the metadata records in an easy way.

4.3.1.2 Persistent Identification of Resources (PIDs)

Each Netherlands CLARIN centre has set up a PID-system for the creation, the assignment, the maintenance and the resolution of persistent identifiers. Initially, any working system of PIDs was allowed, but in a later phase CLARIN required the use of the Handle System: each CLARIN centre in the Netherlands now uses the Handle System for the assignment and resolution of persistent identifiers.[8]

4.3.1.3 Authentication and Authorisation

CLARIN requires the use of SAML2 Identity Management for user identification and authentication. This provides a unified domain where researchers have a single identity and can use Single Sign On (SSO) for all CLARIN services.

To enable this in the Netherlands each CLARIN centre that hosts protected data or services must be a member of the CLARIN Service Provider Federation (SPF) and the Dutch academic and higher education Identity Federation SURFconext. Such legal contracts with national Identity Federations are necessary, because the users must be sure that they are indeed using an approved service, and not some unknown service that might abuse users' personal information or implicitly charge costs to users or their institutes. All CLARIN centres in the Netherlands are also members of the CLARIN SPF, which takes care of inter-federating the different national Identity Federations with regards to CLARIN services, so that Dutch users may access CLARIN services of other European CLARIN centres and vice versa.

To make SAML2-based FIM possible, in addition to joining the federations, every centre has to deploy Shibboleth middleware (or similar software), and ensure that access to the protected data or service always leads to the Shibboleth system: this middleware enables the provisioning

[7] http://www.datasealofapproval.org/en/ (accessed 14 January 2016).

[8] DANS originally only used its own URN:NBN resolver; however the choice of URN:NBN is no longer CLARIN compatible, and only services based on the Handle System technology such as EPIC or DOI/DataCite are now accepted. DANS now also uses DOIs.

and checking of user credentials at the user's home organisation. SURFsara, which hosts the Dutch National Research and Education network (NREN) organisation, offers administration of FIM services through the SURFconext federation; however, this service is by default a national service, connecting only users and services from Dutch organisations. This is too limited in the CLARIN context, which aims to provide access to all European researchers (and even wider). Making FIM available in a European CLARIN context requires some extra configuration and administrative actions. Some centres ran into this problem, and DANS still needs to make final adjustments.

To enable the interfederation function of the CLARIN SPF, the administrators of the identity stores from the user home organisations (e.g. universities and research organisations) need to give permissions for their students and employees to use CLARIN services. To this end, SURFconext made a request to all their members to allow SURFconext to pass through user information to CLARIN service providers.[9] However, the response to this request was minimal: only a few organisations gave permission. In order to improve the situation, a new strategy was followed. Firstly, a documentation package was prepared to explain exactly what was involved and what type of information would be conveyed to CLARIN service providers. It included a letter by the general director of NWO, in which each organisation was requested to grant permission for usage of the CLARIN services. It referred to the original letter sent by SURFconext. It also included a link to a tool to test, after permission was given, whether the technical implementation of this permission actually worked correctly. Secondly, 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 SURFconext, and might more easily 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 described exactly what the request was, and what it involved. 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, all CLARIN-NL partners have given permission to their employees and students to use the CLARIN services.

The same package has been sent to our colleagues in Flanders, and it has been successfully used there as well, since at least some universities in Flanders (e.g. Ghent University and KU Leuven) have access to the CLARIN services through the CLARIN AAI system. This was especially important for the 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.

With respect to authorisation, CLARIN does not make any requirements regarding the use of a specific system or technology. However, CLARIN does require the hosting centre to make all the legal and ethical requirements for using a specific resource or service explicit.[10] Some CLARIN centres in the Netherlands have special provisions to deal with such matters, e.g. The Language Archive unit of the Max Planck Institute for Psycholinguistics (MPI/TLA). For example, one option is to show users a text describing usage conditions, but letting users access the data without reading this text. In a second option, such a text is shown but confirmation by users that they have read the text and agree to it is required.[11] 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 data set, which requires special provisions to protect the privacy of the participants, who come from the (small) sign-language-using community in the

[9] SURFconext, the Dutch Identity Federation has a 'star' architecture so the technical configuration of such permissions can be done solely by SURFconext itself. In other countries with a different Identity Federation architecture, the Identity Providers must make such adaptations themselves.

[10] Currently CLARIN has developed a simple classification system for such licences: see https://www.clarin.eu/content/license-categories

[11] See https://www.eff.org/wp/clicks-bind-ways-users-agree-online-terms-service for clickwrap and browsewrap agreements.

Netherlands). Most 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.

4.3.1.4 Resources in CLARIN-Recommended Formats

Each CLARIN centre makes resources available in a CLARIN-compatible manner, as was described in detail in chapter 2 (sections 2.2 and 2.4.5). A note of concern is that multiple lists of recommended formats are circulating within CLARIN, and as of today no authorative list has been presented yet. This problem and other issues with respect to standardisation are the domain of the CLARIN standards committee.

4.3.1.5 Data Seal of Approval

CLARIN centres must store data and software. To that end, each CLARIN centre has to set up a repository. Different repository systems exist, and CLARIN does not prescribe which system has to be used. Many centres use an open source repository platform, such as the Fedora Commons repository[12] or DSpace, and some have developed their own software, for instance LAMUS/LAT, developed by the MPI (Broeder et al., 2006).

The CLARIN centre registry contains information on the repository systems used by the various CLARIN centres. Currently each Dutch CLARIN centre uses a different system: DANS uses its own EASY system, which is built on Fedora Commons; INL uses DSpace; MPI/TLA uses LAT; and Huygens ING and the Meertens Institute use their own systems.

Some centres use special software so that users can store resources in the repository; for example, MPI/TLA's LAT and the DANS EASY archiving system offer deposition facilities. Storing resources in the repository must be supported by special software, since it is not an easy matter. Typically PIDs are assigned at 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 and it must be added to the resource description, which now can be finalised and gets its own PID. Provisions for legal or ethical access and usage 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 description must be put on a location where it can be harvested by the OAI-PMH protocol.

Each CLARIN centre must ensure the long-term preservation of the resources; CLARIN centres have to make special provisions for this. Sometimes they take care of long-term preservation themselves (e.g. DANS), but most centres outsource it to specialised centres (e.g. the MPI/TLA outsources it to the long-term preservation services of the Max Planck Gesellschaft). In any case, each CLARIN centre must have a clear procedure in place for managing its data and for ensuring long-term preservation, and must work according to this procedure. Each CLARIN centre must be awarded a DSA it is to become a certified CLARIN centre. The DSA guidelines[13] 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 candidate CLARIN centres in the Netherlands have now been awarded the DSA. Initially some Dutch candidate centres perceived the DSA requirements as unclear or too difficult and beyond their ambitions. This was unexpected, based on information from other national CLARIN projects and the fact that the archives of MPI and DANS had already received the DSA. The main

[12] Fedora stands for 'Flexible Extensible Digital Object Repository Architecture'.

[13] See https://www.datasealofapproval.org/en/community/, in particular guidelines 6,7 and 8.

issue turned out to be insufficient clarity as to the extent to which the DSA requirements at that time also allowed centres to outsource the required persistent archiving functions to other organisations. Information meetings on certification topics organised by the CLARIN ERIC did not resolve this issue – this caused some delays. As a positive side effect of this it can be mentioned that in the current version of the DSA it has been made explicit that all guidelines can be outsourced (as long as the outsource partner has a DSA or better level of trust certification).

4.3.1.6 Evaluation of the Certification Process

At the start of the IIP, the full CLARIN certification procedure as it stands now was not yet in place beyond the requirement to obtain the DSA. However, formal certification was always seen as one of the major targets to achieve. Though a lot of preparatory work could be done before, focused activities on certification started only when the requirements were fully clear and these were adopted in the IIP work plan.

The speed and final level of implementation and enabling of the specific CLARIN infrastructure technologies varied across the different candidate centres. This was in part due to their different histories: some of them were already involved in earlier infrastructure projects – e.g. DAM-LR (Broeder et al., 2008) or the EU-funded CLARIN-Preparatory Phase project – and two prospective CLARIN centres had also already acquired the DSA.

The first CLARIN-NL centre that obtained its CLARIN certification was the MPI, as it was also part of the German CLARIN-D project that started the certification process as soon as initial certification procedures were in place. This enabled the MPI to provide advice and hands-on help to the other Dutch centres in their preparations for certification. The other CLARIN-NL centres passed certification at different moments. Challenges encountered varied from problems with configuration requirements for joining the CLARIN SPF to problems regarding the full integration of required services with already existing ones, such as authentication and authorisation. Currently (in December 2015) only DANS is still not fully certified as a type B CLARIN centre. One of the reasons is that DANS needs to satisfy multiple groups of customers outside the direct CLARIN domain and thus has problems committing to CLARIN service requirements, for which it already has existing (CLARIN-incompatible) services in place.

Another concern raised by the candidate centres was the changes of CLARIN certification requirements during the run of the IIP project without clear communication. Although in a developing infrastructure such changes are unavoidable, it became clear that the existing communication channels did not always reach all involved parties, and improvements have been made to communicate all relevant information from the CLARIN committees directly to the (candidate) centres.

4.3.1.7 Stability

The set-up of the CLARIN infrastructure as a network of CLARIN centres is intended to create a flexible and robust infrastructure. During the CLARIN-NL project this network was put to the test on several occasions.

First, the HLT Agency (TST-Centrale) was split off from the INL and turned into the CLARIN centre of the Dutch Language Union. This resulted in a lot of additional work to implement the split and created a situation where the HLT Agency had to start from scratch again in becoming a certified CLARIN centre. The HLT Agency worked towards this goal, but, because of financial problems at the Dutch Language Union, the necessary staff could not be made available, and most available employees started looking for other opportunities, reducing the staff even further. Fortunately, the HLT Agency did create CMDI metadata for the data that it manages, and these can be found via the Virtual Language Observatory (VLO). Recently, the data of the HLT Agency also went back to INL.

At the same time, there were managerial problems at INL, and the future of the INL was reassessed. It was clear that a large restructuring was going to take place. This created a lot of uncertainty among the INL staff, which was not optimal. Fortunately, that period is now over, and INL is turning into the Institute for the Dutch Language (INT) and plays a full role in the CLARIN-NL successor project CLARIAH.

Most problematic, however, was the decision of the MPI management to stop its activities as a CLARIN A-centre, and minimise its activities as a CLARIN B-Centre to the level of what was contractually required by its participation in the German CLARIN-D project. This required several adaptations. We concentrate here on the consequences for the Dutch B-centres (for the consequence related to the type A services, see section 4.3.3). MPI decided that the archiving system and workflow used was too complex for their smaller role, so an activity was set to design and construct a simpler version. This has been done in the context of The Language Archive collaboration (TLA), a consortium of MPI, DANS and the Meertens Institute. It resulted in a new CLARIN-compatible repository system called Fedora Language Archiving Technology (FLAT; Windhouwer et al., 2016). Testing the use of FLAT for depositing the results of some previous CLARIN data curation projects at the Meertens Institute was financed by CLARIN-NL (the MDF project).

In addition, it was concluded that it might be beneficial to have important resources available at multiple centres, and an action was started to ensure that resources hosted by MPI are also hosted by other CLARIN centres in the Netherlands. Nevertheless, the reduction of the activities by MPI also means that there is no natural centre for specific types of language data, especially if they concern languages other than the Dutch language (both INL/INT and the Meertens Institute focus on the Dutch language).

4.3.2 Type D Centres

The missions of resource-providing centres vary, and in many cases such centres need to satisfy multiple groups of customers outside the direct CLARIN domain and thus cannot commit themselves exclusively to CLARIN requirements. To address this issue, CLARIN-NL introduced the NL-specific CLARIN data providers centre type D[14] for centres not exclusively focusing on typical CLARIN resources or services.

CLARIN centres of this special type (called CLARIN-NL Data Providers or Type D CLARIN centres[15]) distribute data independently of CLARIN (and have been doing this long before its establishment), but have made provisions to give access to the data that are relevant to humanities researchers in a CLARIN-compatible manner (via CMDI resource descriptions). These CLARIN centres include organisations that, by their very mission, make available large amounts of data. Currently the Type D centres are:

- Koninklijke Bibliotheek (KB)[16] for digital books, articles, newspapers.
- Digitale Bibliotheek voor de Nederlandse Letteren (DBNL; now included within KB)[17] for literary works.
- Nederlands Instituut voor Beeld & Geluid (NIBG)[18] for audio-visual data (especially TV and radio programmes).
- Utrechtse Universiteitsbibliotheek (UBU)[19] for digital books and articles.

[14] See http://www.clarin.nl/sites/default/files/restore/New%20Centre%20Types%20110607.pdf and http://www.clarin.nl/node/130 (accessed 14 January 2016).

[15] This type of CLARIN centre is distinguished from others only in the Netherlands.

[16] National Library.

[17] Digital Library for Dutch Literature, which merged with KB in 2014.

[18] Netherlands Institute for Sound and Vision.

[19] Utrecht University Library.

Since many of the data provided by these organisations are highly relevant to humanities researchers, these data should be available via the CLARIN infrastructure – and they already are, for NIBG and for UBU, or will soon be (for KB).[20]

Although the CLARIN Data Providers were subsidised in separate CLARIN subprojects, the IIP supported these efforts, e.g. with the development of suitable CMDI metadata schema. Such smaller consultancy services were also rendered by the IIP to the CLARIN-NL curation and demonstrator projects and to other CLARIN-NL projects, e.g. the DCS.

4.3.3 Type A Centres

Type A centres offer core, essential infrastructure services; for example, the MPI/TLA hosts[21] the VLO (see section 4.6.3), and carries out the necessary harvesting of resource descriptions. A list of essential services is available via CLARIN.[22]

Only the MPI/TLA and the Meertens Institute offer type A services (in addition to the type B services they offer) in the CLARIN-NL project. MPI/TLA offer many type A services, and the Meertens Institute a few; for example, the Meertens Institute hosts the CLARIN Concept Registry (CCR; see section 4.6.2) and CLAVAS (see chapter 5).

CLARIN-NL contributed to these core infrastructural services, which was quite natural since MPI was involved in building and managing many of these services. CLARIN-NL also contributed significantly to registries and services for CMDI metadata, e.g. the CMDI registry, the Arbil metadata editor, the VLO and the CLARIN discovery services (see section 4.6).

An official certification procedure for type A services only became available in 2016, long after the end of the IIP; but the concept of a CLARIN type A centre existed since the start of CLARIN, and CLARIN-NL, through the IIP, contributed considerably to the maintenance and further development of these core services.

4.4 IIP Strategy for Technical Infrastructure Adoption by Centres

4.4.1 Knowledge Exchange and Consultancy

To build a technical infrastructure as used by CLARIN, it is crucial that the different centres have expert staff available and that there be lines of communication and information and training possibilities to keep their knowledge up-to-date.

During the run of the IIP, many meetings and workshops were held to discuss technical infrastructure aspects necessary for CLARIN certification and interoperability with the CLARIN infrastructure. Prior to the certification round at the end of 2013, direct consultancy by the previously certified[23] MPI CLARIN centre was provided to other centres, as also happened for the second round in April 2014.

4.4.2 CLARIN-NL Infrastructure Work in a European and Historical Context

Both MPI and INL had been involved in the DAM-LR EU project, which pioneered technical solutions that would become part and parcel of the CLARIN infrastructure, such as SAML-based

[20] The number of metadata records available from KB is so huge that it requires modifying the VLO.

[21] Because of the MPI's changed CLARIN ambitions the VLO service has been moved and is currenty under the direct responsibility of the CLARIN ERIC.

[22] See http://www.clarin.eu/content/services for a list of such essential infrastructure services.

[23] As mentioned earlier, MPI had already been certified in the certification round for the centres in the German CLARIN project.

FIM and PIDs based on the Handle System (HS). Within INL, the DAM-LR work was carried out by the Dutch-Flemish HLT Agency, then hosted at the INL. Both INL and MPI had an operational Handle System server installation and knew about the effort required for administrating this. From the earlier experiences with SAML-based FIM a concern arose over the increasing complexity of the managed metadata and the policing of the authentication system for every provided service. This was solved by centralising the SAML2 service on a web server that acts as a 'reverse proxy' and where all Shibboleth configuration takes place and then publishing all to-be-protected services via this proxy. This solution was implemented both at MPI and INL.

CLARIN requires the use of CMDI metadata for Language Resources and Language Technology services, which differs from the approach to metadata in DAM-LR. IIP supported the CMDI development, which was already initiated during the CLARIN EU preparatory phase (CLARIN-PP), and had already resulted in creating the necessary software components at the MPI (see section 4.6). The applicability of the CMDI approach to a variety of resources stored in the candidate CLARIN centres was first tested in a small project (De Vriend et al., 2013).

As a PID solution, most CLARIN centres chose, based on their experiences in DAM-LR, either to provide their own instance of the Handle System or to use Handle System PID services provided by EPIC (this was the case the Meertens Institute, for instance). In the Netherlands, EPIC handle services are delivered by SURFsara.[24]

To realise the CLARIN AAI objectives, it was necessary for all CLARIN-NL centres to join the CLARIN EU SPF, which, as mentioned before, implied that the centres also join the Dutch Identity Federation SURFconext. The MPI provided support to adopt the necessary technological knowledge and coordination with SURFconext and other EU AAI organisations. Although current AAI solutions provide satisfactory authentication and authorisation procedures for accessing web-applications, the necessary AAI infrastructure for the orchestration of web-services was only tested on a small scale and never fully deployed (Blumtritt et al., 2015). This hampered the further development of more advanced use cases where services authenticate on behalf of the user to perform advanced types of analysis.

With respect to the development of a flexible workflow system for Language Technology, no convergence was achieved at the CLARIN EU level and, as mentioned above, a suitable stable AAI technology for such a system was lacking, although prototypes were successfully tested. The CLARIN-NL TTNWW project delivered a functional workflow system based on a number of fixed recipes (see chapter 7).

With respect to LR format standards especially, the IIP based itself on the work done in the CLARIN EU preparatory phase, extending the accepted set of CLARIN-allowed data formats with the new 'Folia' annotation format used widely in the Low Countries (see Van Gompel et al., 2013; and chapter 6).

4.5 Populating the CLARIN Infrastructure

With centre certification, a second aspect of the centre build-up activities has been making the centres' data accessible and available in the CLARIN infrastructure. Partly this was covered by the requirement that all results of (other) CLARIN projects should be provided in a CLARIN-compatible way, but it was also expected that, as part of the IIP, centres would transform already existing data sets into CLARIN-compatible ones.

[24] See https://www.surf.nl/en/services-and-products/data-persistent-identifier/index.html (accessed 14 January 2016).

Continuous work was performed by the candidate centres to make their resources available. The result of this work can be found on the CLARIN-NL portal,[25] which shows all the corpora and services that were created in CLARIN-NL and that can now be accessed in a CLARIN-compatible manner, or by using the VLO CLARIN metadata catalogue[26] or CLARIN centre portal pages such as the INL CLARIN Portal.[27]

A complication in this process for the INL centre was the split-off of their TST-centrale (Dutch-Flemish HLT Agency) partner, which is currently the candidate CLARIN centre for the Dutch Language Union. Also, although this occurred after the official end of the IIP, the MPI centre decided to scale down its efforts in infrastructure building and provisioning projects. Such events prove very challenging for infrastructures such as CLARIN that depend on centres as stable components of the infrastructure for providing essential data and services. With respect to the provisioning of services, we have been able to partially overcome such changes by involving other centres or moving services to other centres or to CLARIN ERIC. With respect to data provisioning, sustainability very much depends on the possibility to make copies available – this is not always easily possible in view of copyright and licence issues.

4.6 Contributions to Central CLARIN Registries and Services

An important requirement for a functioning CLARIN infrastructure is the availability of certain specific *central* registries and services. The IIP made important contributions to this. These registries and services mainly concern the CLARIN Component Metadata Infrastructure registry and registries for managing semantics, as well as registries used in CLARIN data and central services that are needed for using FIM for authentication to CLARIN services. The European scope requires that these services and registries also be available EU-wide, and they should thus be considered a major contribution to the EU CLARIN infrastructure.

In the IIP, MPI was assigned the task of developing and maintaining the registries and services. MPI had already started developing prototypes within the CLARIN EU preparatory phase and was well positioned to coordinate with the other national projects, especially CLARIN-D, in which it also participated. The German CLARIN project also provided considerable support to the development and maintenance of these registries and services.

The relevant registries and services are:

- *The CMDI Component Registry and Editor*. A web-based tool[28] that allows researchers and data scientists to reuse and create new (CMDI) metadata profiles to create metadata records.
- *The ISOcat Data Category Registry (DCR) and later the CLARIN Concept Registry (CCR)*. Registries providing semantic interoperability for concepts used within metadata and annotations.
- *VLO, the Virtual Language Observatory*. A faceted browser that shows all the metadata records available within the CMDI domain. This service includes the metadata-harvesting infrastructure.
- *Arbil*[29]. A metadata editor that users can use to create CMDI metadata records
- *The CLARIN Discovery Service and the CLARIN Identity Provider for users without an academic institute affiliation (homeless).*

[25] The CLARIN NL portal can be found at: http://portal.clarin.nl/clarin-data-list-fs

[26] https://vlo.clarin.eu/search?8& fq=nationalProject:CLARIN-NL

[27] https://portal.clarin.inl.nl

[28] https://www.clarin.eu/cmdi

[29] http://tla.mpi.nl/tools/tla-tools/arbil/

Of these we will discuss the CMDI Component Registry; ISOcat and the CCR; and the VLO in more detail.

4.6.1 CMDI Component Registry

Apart from the continuous support and improvements to the stability and performance of the CMDI Component Registry, its functionality was also extended in the IIP. Since Component Registry content management is required not only for CLARIN-NL but also for other CLARIN EU projects, coordination and close collaboration was required. A matter of concern is that projects occasionally choose to create or modify their metadata schema outside of the Component Registry. This can lead to incompatible metadata, a matter that is not always noticed immediately since the central metadata-harvesting process does not check the records thoroughly enough. This is currently addressed by the CLARIN ERIC CMDI and the Metadata Curation task forces.

Another matter of concern connected to the metadata records produced for CLARIN is the sometimes questionable quality of these records, both in semantic correctness and in suitability of the created schema for a particular type of data. This is not a matter that can be solved by technological means, but one that needs to be addressed by curation of the schema and community review. Having metadata content quality managers, as is planned, will be a good step forward.[30]

4.6.2 ISOcat and the CCR

In 2009 the ISOcat DCR (Broeder et al., 2014) went into production as a joint effort between ISO TC 37 and CLARIN. Within the IIP, ISOcat has been extended to better fit into the developing CLARIN infrastructure. The most salient additions are the implementation of FIM for ISOcat, the interoperability with the CLARIN CLAVAS vocabulary service (see chapter 5) and the supporting of community recommendations gathered by the CLARIN ISOcat content manager. In 2013 there was an extensive evaluation of the problems encountered and of both ISO TC 37's and CLARIN's uptake of ISOcat. In a joint meeting of these two major user communities at the end of 2013, the extent of overlap of the requirements of both communities was further assessed. As a result CLARIN and ISO TC 37 developments were further decoupled. It was decided that the design and implementation of a successor to ISOcat would need to be based on a simpler data model focused on concept specifications and on a workflow that is more geared towards community agreement, rather than towards an official ISO standardisation process ISO 12620. Since the MPI CLARIN centre has also expressed its desire to lessen its responsibility towards external infrastructures, CLARIN has switched to the CLARIN Concept Registry (CCR), which is based on OpenSKOS[31] and hosted by the Meertens Institute. A successful migration at the start of 2015 proved to be a showcase of a successful transfer of software, data and responsibilities between centres.

The IIP supported the ISOcat coordinator and various CLARIN projects through tutorials, workshops, question-answering and help with regard to the import/export of data category specifications.

4.6.3 VLO

The Virtual Language Observatory (Van Uytvanck et al., 2012) is the CLARIN EU-wide metadata catalogue that harvests all CLARIN-compatible metadata. The IIP contributed to the development of this application, which serves the whole CLARIN EU community. Much effort was made to make

[30] In the new CLARIAH-core project extra efforts will also be spent on metadata quality issues.

[31] http://openskos.org/

the VLO more stable and to make it work with the semantic mappings provided by the ISOcat and CCR registries. Related to the VLO is the metadata-harvesting process itself, which occasionally requires intensive communication with the metadata providers when problems occur.

4.6.4 Further CLARIN EU Collaboration

The IIP also contributed to the CLARIN EU task forces and committees: these are CLARIN ERIC coordinated groups to discuss and further the different CLARIN infrastructure components. The IIP contributed to the CLARIN metadata task force, the CLARIN AAI task force, the CLARIN Standards Committee and the CLARIN Centre Assessment Committee.

4.7 Conclusions and Recommendations

In this section we describe achievements and hopeful beginnings, but also some lessons learned that should be taken up by future projects such as the continuation of CLARIN-NL within the more broadly scoped CLARIAH CORE project.

First of all, four CLARIN centres were certified during the CLARIN-NL project, and the centres built up a critical mass of expertise enabling them to participate in the next phases of (also European) infrastructure construction and exploitation. At the beginning of the project, most of this expertise was only available at MPI and INL, but this was successfully transferred to the new CLARIN centres, and, additionally, in discussions with the new CLARIN D-type centres we exchanged expertise with library partners (such as KB and UBU) and Cultural Heritage institutes (such as Sound and Vision).

Secondly, although the direct funding came from outside the IIP, the IIP has succeeded in offering opportunities for a wide range of humanities researchers to familiarise themselves with typical research infrastructure services such as aggregated metadata catalogues, persistent identifiers etc.

The CLARIN-NL project was also very successful since it enabled all centres involved to gain knowledge and expertise about the different aspects of research infrastructures and the technical and organisational challenges involved in providing data and services in new ways not always immediately related to a centre's own research groups or customer base.

On an organisational level we have seen that the need for sustainability cannot be addressed by relying on the CLARIN centre model only. The model has to be augmented by some capable central agency able to step in (if even temporarily) and take care of essential central services; alternatively, suitable incentives for CLARIN centres to take over orphaned services must be put in place. Of course, such services must be built from the beginning as 'transferable' services, e.g. with no hidden local dependencies and built on open software. The new requirements for CLARIN A-type services cover such aspects. During and just after the IIP we saw that the performance of two Dutch CLARIN centres (MPI and INL) was compromised by the changing goals of their organisations. Fortunately, it was possible to transfer essential services largely to the direct care of CLARIN ERIC and to the Meertens Institute, but it proved very difficult to find CLARIN centres (even outside the Netherlands) prepared to take over responsibility for such services. In this respect, the CLARIN-NL landscape proved vulnerable.

An interesting issue is the relative difficulty with which the Dutch candidate CLARIN centres were certified compared to the experiences in the German project. This can maybe partly be explained by the fact that the German centres initially[32] were certified earlier (with somewhat lighter requirements). However, in the Netherlands the motivation to adopt CLARIN requirements

[32] However, in a new (re)certification wave three years after the initial certification, with more heavier requirements, the German centres also proved successful.

often seemed less pronounced and the discussions to reach a consensus and to agree to implement CLARIN centre requirements took much time and effort. Since the candidate centres all agreed to the CLARIN goals, this could be an issue of internal institute strategy alignment, or of a need for a more specific CLARIN argumentation at all necessary organisational levels in the Dutch context. But we should also add that the CLARIN centre requirements were under development, and it was felt that this development process was outside the direct influence of those involved. In this respect, increased participation and the need to foster more support for CLARIN requirements and policies from those needed to implement and work with them is a strong recommendation.

In the case of centres with a more heterogeneous mission, it is less certain that all CLARIN centre requirements can be taken up with sufficient priority. The CLARIN-NL project planning itself was targeted towards centre organisations mainly concerned with language data, where the CLARIN scope aligns with a large part of the existing activities. For centres with a broader or more heterogeneous mission, however, special road maps may be needed if their CLARIN participation at a high level is required or desired. It is matter of political consideration how much effort and resources should be invested. A different CLARIN centre classification, such as C, K or D, should also be considered in such cases.

References

J. Blumtritt, W. Elbers, M. Hinrichs, W. Qiu, T. Goosen, M. Sallé, M. Windhouwer. User Delegation in the CLARIN Infrastructure. In J. Odijk (ed.), *Selected Papers from the CLARIN 2014 Conference*. Linköping Electronic Conference Proceedings, August, 2015.

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).

D. Broeder, D. Nathan, S. Strömqvist, S., & R. van Veenendaal. Building a federation of Language Resource Repositories; the DAM-LR project and its continuation within CLARIN. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), European Language Resources Association (ELRA), Marrakech, Morocco, May 28–30, 2008.

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., & Wittenburg, P. (2006). LAMUS: The Language Archive Management and Upload System. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006) (pp. 2291–2294).

D. Broeder, I. Schuurman, M. Windhouwer. Experiences with the ISOcat Data Category Registry. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014), European Language Resources Association (ELRA), Reykjavik, Iceland, May 28–30, 2014.

F. de Vriend, D. Broeder, G. Depoorter, L. van Eerten, D. van Uytvanck. Creating & testing CLARIN metadata components. *Language Resources and Evaluation*. 2013. Vol 47, 1315–1326.

ISO 12620:2009. *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources*. International Organization for Standardization, Geneve, Switzerland, December, 2009.

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

M. van Gompel, M. Reynaert. Folia: A practical XML Format for Linguistic Annotation – a descriptive and comparative study. In Computational Linguistics in the Netherlands Journal 3, 2013.

D. van Uytvanck, H. Stehouwer, L. Lampen. Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC 2012), European Language Resources Association (ELRA), Istanbul, Turkey, May 23–25, 2012.

T. Váradi, S. Krauwer, P. Wittenburg, M. Wynn, K. Koskenniemi. CLARIN: Common Language Resource and Technology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), European Language Resources Association (ELRA), Marrakech, Morocco, May 28–30, 2008.

M. Windhouwer, M. Kemps-Snijders, P. Trilsbeek, A. Moreira, B. van der Veen, G. Silva, D. von Rhein. FLAT: constructing a CLARIN compatible home for language resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), European Language Resources Association (ELRA), Portorož, Slovenia, May 23–28, 2016.

J. Zhang, M. Kemps-Snijders, H. Bennis. The CMDI MI Search Engine: Access to language resources and tools using heterogeneous metadata schemas. *Theory and Practice of Digital Libraries*. Springer Berlin Heidelberg, 2012. 492–495.