Recent Developments in CLARIN-NL

Jan Odijk

UiL-OTS, Utrecht University
Trans 10 3512 JK Utrecht, The Netherlands
j.odijk@uu.nl

Abstract

In this paper we describe recent developments in the CLARIN-NL project with the goal of sharing information on and experiences in this project with the community outside of the Netherlands. We discuss a variety of subprojects to actually implement the infrastructure, to provide functionality for search in metadata and the actual data, resource curation and demonstration projects, the Data Curation Service, actions to improve semantic interoperability and coordinate work on it, involvement of CLARIN Data Providers, education and training, outreach activities, and cooperation with other projects. Based on these experiences, we provide some recommendations for related projects. The recommendations concern a variety of topics including the organisation of an infrastructure project as a function of the types of tasks that have to be carried out, involvement of the targeted users, metadata, semantic interoperability and the role of registries, measures to maximally ensure sustainability, and cooperation with similar projects in other countries.

Keywords: research infrastructure, data centres, resource curation

1. Introduction

In this paper we describe recent developments in the CLARIN-NL project[1] with the goal of sharing information on and experiences in this project with the community outside of the Netherlands and provide some recommendations for related projects, e.g. other national CLARIN projects. It is a natural sequel to the LREC 2010 contribution (Odijk, 2010).

We first describe the basic intended functionality of the CLARIN Infrastructure, and the requirements that it must meet to realize this functionality (section 2.). In section 3., we describe the CLARIN-NL project. In section 4. we describe recent developments in CLARIN-NL. The experiences in CLARIN-NL lead to a number of recommendations summarized in section 5.. Finally, our conclusions can be found in section 6..

2. The CLARIN Infrastructure

CLARIN-NL aims to set up the Netherlands part of the CLARIN infrastructure, a technical research infrastructure for humanities researchers who work with language data. The researcher should easily find the data needed for his/her research in the CLARIN infrastructure, and this infrastructure should have to be the only place he/she has to look for data. The researcher should also easily find software tools needed for his/her research in this infrastructure. The infrastructure should enable the researcher to apply this software to the research data in a seamless manner (without any ad-hoc adaptations of the software or the data). And the researcher must be able to put any new or enriched data or software tools created in the research project into the infrastructure with little effort, so that they become available to other researchers for their research, for creating opportunities for replicating and verifying reported research results, and for creating enriched publications.

Designing and implementing an infrastructure as described above entails a lot of requirements, among them:

- Existing data and tools must be made visible and accessible using standardized metadata. Metadata registries must, via a portal, offer facilities for browsing, searching and selection of data and metadata.

- Existing metadata, data and tools must be made interoperable, not only on the formal level via standards and best practices for formats but also on the semantic level via explicit semantics encoded in a standardized way using commonly agreed upon data categories stored in data category registries, relations between data categories, etc.

- Facilities must be offered to apply software to data in a seamless manner, e.g. as web services in a work flow system. New or enriched data derived from applying software to data must be enriched with metadata including provenance information in a maximally automatic manner.

- Data, tools and metadata must be referable in a persistent way, e.g. by using persistent identifiers, and they must be stored in a way that ensures long term preservation.

- Facilities must be offered to process large volumes of data, and to execute computationally intensive computations.

In this paper we report major aspects of what CLARIN-NL has done since its start in April 2009 to contribute to realizing these requirements.

3. CLARIN-NL

CLARIN-NL forms the Netherlands national counterpart of the CLARIN enterprise on the European level (CLARIN-EU). It complements the preparatory project at the European level (CLARIN-prep)[2] that finished Mid 2011,

---

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

[2] http://www.clarin.eu/

and is one of the national projects that will contribute to designing, implementing, and exploiting the Europewide CLARIN infrastructure that will be managed by the CLARIN ERIC, which has been established on Feb 29, 2012.

CLARIN-NL has been set up as a mixture between a programme and a project. Its nature as a programme with multiple calls for subprojects spread out over the duration of the CLARIN-NL project provides the flexibility to adapt the contents to new developments, and to new players in the field (e.g. humanities researchers not reached yet). It also increases the involvement of a maximally large research group (since the money has not all been distributed already over the initial partners). Its nature as a project offers opportunities for defining a few tightly organized longer term subprojects in selected areas with selected expert groups so that knowledge and expertise built up will be sustained in the participating institutes.

The CLARIN-NL proposal covers a period of 6 years, partitioned in three phases of two years: the preparation phase, the construction phase, and the first two years of the exploitation phase. CLARIN-NL has effectively started on April 1, 2009 and will run for six years (2009-2014). Its budget is 9.01 million euro. CLARIN-NL is the first CLARIN-related national project that has been awarded money not only for the preparatory phase but also for the implementation and exploitation phases. We submit that our plans and experiences can serve as an example that other national CLARIN projects and project proposals can profit from.

Currently CLARIN-NL has 26 participants from linguistics and the humanities more broadly, and the number of participants is growing. The participants include all universities in the Netherlands with a humanities faculty, several university institutes, the humanities institutes of the KNAW (Royal Academy), independent institutes, libraries, and data centres.[3]

4. Recent Development

We report on some of most important current developments in CLARIN-NL.

4.1. Infrastructure Implementation

Within CLARIN-NL, a range of infrastructure implementation subprojects have been started up, including:

IIP A subproject *Infrastructure Implementation Project* (IIP) for implementing basic infrastructural facilities, such as setting up:

- a federation of Dutch CLARIN Centres

- facilities for metadata creation, registry and harvesting

- facilities for semantic interoperability (ISOCAT, RELCAT, SchemaCAT)[4]

- facilities for assigning and resolving persistent identifiers (PIDs)

- orchestration of web services in a workflow system

- an authentication and authorization infrastructure for providing single sign-on facilities for all CLARIN services

- a registry for obligatory and recommended CLARIN formats

- relations with large data and computational centres such as SARA[5], and the Netherlands eScience Centre[6]

Participants in the IIP project are the centres that expressed the ambition and made the commitment to become a CLARIN type A/B centre in the Netherlands, i.e. MPI[7] (Nijmegen), Meertens Institute[8] (Amsterdam), INL[9] (Leiden), and DANS[10] (The Hague), and more recently Huygens ING[11] (The Hague).

MD A *Metadata* (MD) subproject for testing the Component Metadata Infrastructure (CMDI)[12] against data residing in Dutch research organizations. In this subprojects metadata components and profiles to describe data were created and tested against existing resources in the Netherlands by experts. This subproject has restricted itself to metadata for data, but components and profiles for metadata for software (command line tools, applications, web services, etc.) are also desperately needed. A follow-up project to do this is to start this year.

S&D A *Search & Develop* (S&D) subproject for providing intelligent search facilities in metadata and in the actual data. The S&D project aims to create a network of centres in the Netherlands, focusing around search services. The S&D project aims to create search facilities for search in metadata stored in a centralized database (itself filled and regularly updated via metadata harvesting), as well as federated search facilities in the actual data. A first prototype S&D federated search demonstrator with limited functionality (term search only) has been created.[13] Researchers have been requested to submit user scenarios, but the few that have been submitted[14] are too complex for the initial prototypes and will be reconsidered in a later phase of the project. Though the S&D project is running according to plan, it has become clear that certain highly resource-specific content search engines, e.g. the ones for EuroWordNet (Vossen, 1998) and Cornetto[15], and DACT[16] as used in LASSY[17], cannot be included as part of this

---

[3]See http://www.clarin.nl/node/2, under *Participants* for a list of the participants

[4]http://www.isocat.org and (Kemps-Snijders et al., 2010).

[5]http://www.sara.nl

[6]http://www.esciencecenter.com/

[7]http://www.mpi.nl

[8]http://www.meertens.knaw.nl

[9]http://www.inl.nl

[10]http://www.dans.knaw.nl

[11]]http://www.huygens.knaw.nl/

[12]http://www.clarin.eu/cmdi and (Broeder et al., 2010).

[13]http://lux17.mpi.nl/ds/fedsearch/

[14]http://www.clarin.nl/node/218

[15]http://www.inl.nl/tst-centrale/nl/producten/lexica/cornetto/7-56

[16]http://rug-compling.github.com/dact/

[17]http://www.let.rug.nl/vannoord/Lassy/

project. Since SRU/CQL (Contextual Query Language)[18] is used in federated search, granularity of access that can be provided by participating content search engines depends largely on the level of CQL support that native search engines are able and willing to provide (mapping onto CQL) and the expressive power of CQL in relation to native implementations (CQL abilities). The CLARIN infrastructure will therefore have to support these content search engines as stand-alone applications.

CLAVAS ISOCAT (Kemps-Snijders et al., 2010) is the preferred data category registry in CLARIN. However, it has become clear that ISOCAT cannot function as the only concept registry in the CLARIN infrastructure. For certain concepts ISOCAT is not particularly suited, and for other concepts independently maintained registries exist and duplication is to be avoided. For this reason, a separate project, CLAVAS (CLARIN Vocabulary Access Service), was started up to build an interface to a variety of concept and data category registries, so that CLARIN users can transparently use concepts and data categories from each of them. For example, standardized codes for languages[19] and countries[20] are maintained by ISO in independent databases, and it makes no sense to copy these into ISOCAT. ISOCAT is also not particularly suited for hosting names of organizations and institutions, and it does not offer functionality to map variants of these names onto the canonical name, though this is highly desirable for data categories such as titles, organization names, etc. CLAVAS will provide such functionality.

Integration A lot of work has been done in the context of CLARIN since the start of the CLARIN-NL project in 2009, in building bricks and pieces of the CLARIN infrastructure, but all of these activities were distributed and have been done in isolation. It is now time to start working towards an integration of all the work done in a real and visible CLARIN infrastructure, and also to organize the remaining work around this CLARIN infrastructure. Integrating all the functionality and data will also test the interoperability and will help us identify functionality that is still lacking and should be added. The initial version of a plan for this integration plans a web site that will serve as a user-friendly portal to the (Netherlands part of the) CLARIN infrastructure, and will make all infrastructure functionality that has already been implemented, as well as all curated resources and demonstrators available to the targeted CLARIN users (via their metadata). The portal will after that also function as scaffold on which to implement extensions of the CLARIN infrastructure. The implementation of this integration platform will be accompanied by intensive PR, educational and training activities to inform the users of the presence of new functionality or resources, to educate them about their nature, and to train them in working with these new components in the CLARIN infrastructure. Undoubtedly, we will encounter problems while carrying out

this integration, but these integration activities start early enough so that lacking functionality, or incompatibility between functionality can still be dealt with. The integration plan is currently being elaborated and its implementation is expected to start Mid 2012.

4.2. Resource Curation and Demonstrator Projects

The goal of a **resource curation** project is to adapt an existing resource so that it is visible, uniquely referable and accessible via the web, and properly documented.

The goal of a **demonstrator** project is to create a documented web application starting from an existing tool or application that can be used as a demonstrator and function as a showcase of the type of functionality CLARIN will incorporate and support.

Important goals common to both types of projects are (1) apply standards and best practices[21] and make use of the suggested CLARIN architecture and agreements to understand their limitations and the requirements for extensions; and (2) establish requirements and desiderata for the CLARIN infrastructure.

In particular, all projects will have to make metadata for their resources in accordance with CMDI, and contribute to semantic interoperability by mapping the data categories used to data categories in ISOCAT, or by creating new data categories in ISOCAT.

In this way, curated resources and a range of showcases of CLARIN functionality will become available, and at the same time evidence-based requirements and desiderata for the CLARIN infrastructure and supported standards and best practices will be obtained. This will make it possible to influence the final selection of standards and best practices that will be promoted by CLARIN.

In addition, each project has to involve intended users of the CLARIN infrastructure, which contributes to bringing the communities of the humanities researchers and technology providers (infrastructure specialists and language and speech technology researchers) together in concrete projects in which they are cooperating. This is essential, since only in this way can there be guarantees that the infrastructure will provide the functionality that is actually needed by the researchers.

Three calls for resource curation and demonstrator projects have been launched (in 2009, 2010 and 2011). Projects from the First Call[22] have finished. Projects from the Second Call[23] have finished or are in their final stages. Third Call projects started early 2012.

The first and the second call were Open Calls. in an Open Call any humanities researcher in the Netherlands can submit a proposal and the project proposals compete among each other for funding. The Third Call, however, consisted of an Open Call and Closed Call. A researcher can submit a project proposal in the Closed Call only upon invitation. The proposals submitted in the closed call are evaluated for quality and relevance but they do not compete with other

---

[18]http://www.loc.gov/standards/sru/specs/cql.html

[19]ISO 639, http://en.wikipedia.org/wiki/ISO_639.

[20]ISO 3166, http://www.iso.org/iso/country_codes

[21]See http://www.clarin.eu/recommendations for a list of standards and best practices currently promoted by CLARIN.

[22]http://www.clarin.nl/node/70

[23]http://www.clarin.nl/node/162

project proposals. Five disciplines that were underrepresented so far in CLARIN-NL were selected for the Closed Call, and for each discipline one researcher was invited to make a project proposal.

The Closed Call resulted in 4 projects from the disciplines *philology / classical studies*, *language acquisition and language impairment*, *religion studies*, and *discourse studies*. For the discipline *descriptive linguistics / language documentation*, no proposal was submitted.

The Open Call resulted in 9 awarded projects on topics from linguistics, history and literary studies.

More details about the CLARIN-NL resource curation and demonstrator projects can be found on the CLARIN-NL web site.[24]

4.3. Data Curation Service

A plan for a Data Curation Service (DCS)[25] was developed. After its approval, the DCS has started preparatory activities in the fall of 2011 and actual data curation as of 2012. The initial data set to be curated is the Dutch Bilingualism Database (DBD). While curating these data, the DCS will investigate, plan and budget for additional candidates to be curated.

The DCS was set up because it will allow us to create a team of experts in the area of data curation, which can profit from the fact that they curate multiple resources. In this way they can extend the necessary expertise, create synergy, and carry out data curation services in a more efficient manner.

4.4. Cooperation with Flanders

A cooperation project between the Netherlands and Flanders (Belgium), especially concerning data and tools specific to the Dutch language (shared by the Netherlands and Flanders), has been set up. In its major subproject, called TTNWW[26], existing language and speech technology tools and data created in the Dutch-Flemish STEVIN project are being made seamlessly interoperable in the form of web services organized in a work flow system. In the context of the TTNWW project, CLAM (Computational Linguistics Application Mediator) has been developed: software to facilitate the conversion of existing tools into RESTful web services[27]. CLAM is successful and is in use by most participants in the TTNWW project.

Furthermore, the corpus format in use in this project, FoliA (van Gompel, 2012)[28], appears to develop into a *de facto* standard for text corpora annotation for Dutch.

4.5. Semantic Interoperability

Actually trying to achieve semantic interoperability via shared data categories is (as could be expected) a difficult task, which raises many questions, some of a principled nature but also very many of a practical nature. Coordination

for this is required to organize activities, to give training and education, to avoid misunderstandings, to make pragmatic decisions taken well-known and justify them, to inventory problems, and to disseminate best practices and practical experiences, etc.

This is especially pertinent since work on ISOCAT to achieve semantic interoperability was carried out in an isolated manner in the various subprojects: assistance was clearly needed, and for many problems a uniform approach should be selected for all subprojects. For this reason, an ISOCAT Coordinator has been appointed to coordinate work on semantic interoperability in the various CLARIN-NL subprojects.

4.6. CLARIN Data Providers

The Netherlands has a range of centres that already for a long time or since recently have making data available in their core mission and that are of great importance to humanities researchers. These include the *Koninklijke Bibliotheek* (KB, Royal Library)[29], *het Instituut voor Beeld & Geluid* (IBG, Institute for Sound and Vision)[30], *het Nationaal Archief* (NA, the National Archive)[31], *de Digitale Bibliotheek van de Nederlandse Literatuur* (DBNL, Digital Library of Dutch Literature)[32], the various university libraries, and some others. CLARIN-NL wants to make sure that data from such organizations are available to humanities researchers in the CLARIN infrastructure and in a CLARIN-compliant manner. For this reason, a special type of centre has been defined, a *CLARIN Data Provider*, and the organizations mentioned have been invited to make a proposal to become such a CLARIN Data Provider. The proposals submitted by KB and IBG have been approved and they are carrying them out. DBNL and NA are preparing their project proposals.

Such an organisation can become a CLARIN Data Provider by meeting a number of requirements[33], the most important of which are:

- It makes metadata available for its data in CMDI format, with the metadata contents in accordance with its own conventions, but with Dublin Core as a minimum;

- It allows harvesting of these metadata by other CLARIN centres in accordance with the protocols required by CLARIN (OAI-PMH);

- It guarantees stable URIs for the resources augmented with a system of unique identifiers that can be used for creating citations.

In this way the CLARIN Data Providers ensure the presence, visibility and accessibility of their data in the CLARIN infrastructure. By the very nature of these organisations, incorporating them in the CLARIN infrastructure is also an excellent way to ensure the sustainability of the

---

[24]http://www.clarin.nl/node/282

[25]http://www.clarin.nl/node/147

[26]An acronym for the Dutch project name *TST Tools voor het Nederlands als Webservices in Workflows*, i.e 'Language and Speech Technology Tools for Dutch as web services in workflows'

[27]http://ilk.uvt.nl/clam/ and (van Gompel et al., 2011)

[28]See also http://ilk.uvt.nl/folia/

---

[29]http://www.kb.nl

[30]http://instituut.beeldengeluid.nl/

[31]http://www.nationaalarchief.nl/

[32]http://www.dbnl.org/

[33]detailed in http://www.clarin.nl/system/files/New%20Centre%20Types%2020110607.pdf

infrastructure, also for the period after which the CLARIN-NL project will have finished.

4.7. Educational and Training Activities

A range of information sessions, tutorials and workshops have been organized, especially on topics that are relatively new to most researchers but crucial for the resource curation process: metadata, and semantic interoperability (ISO-CAT).[34]

A Helpdesk has been set up. It includes a FAQ-section with answers to the most commonly asked questions.[35] The FAQ section is largely in English, so it can also be used by researchers from other countries. CLARIN-NL and CLARIN-D already agreed to cooperate on the Helpdesk.

4.8. Dissemination and Outreach

CLARIN-NL has organized a range of workshops, and it stimulates and supports the organization of workshops by its subprojects and others, especially if they contribute to involving more humanities researchers in the CLARIN infrastructure.

CLARIN-NL has also commissioned the production of a range of short movies to explain and illustrate the work being done in various CLARIN-NL subprojects, with the hope aim they will inspire others to work in, benefit from and contribute to the CLARIN infrastructure. preliminary versions have been shown at various CLARIN-NL events. The movies will be accessible via the CLARIN-NL web site soon.

Several CLARIN-NL subprojects also have made movies or screen captures to illustrate their work or their demonstrators.[36] The impact that these movies will have has still to be determined.

The CLARIN-NL web site has been completely revised and split up in a part intended for dissemination and outreach (reaching people from outside the project) and a part for project-internal matters.

4.9. Cooperation with other projects

CLARIN-NL is closely cooperating with other projects. This is evident for CLARIN Flanders[37] with which we cooperate, inter alia in the TTNWW project mentioned above (section 4.4.).

In a meeting with the German CLARIN-D[38] project team, several topics for cooperation were identified that are currently being elaborated. CLARIN-NL intends to investigate cooperation opportunities with other countries with running national CLARIN projects (e.g. Norway, Austria, and the Czech Republic) as well, and this will surely also be stimulated by the CLARIN ERIC that is now in place and starting its activities. However, it has turned out to be problematic to set up cooperation, since all the funding for many other national projects has already been allocated. It is very desirable that each national CLARIN project has some

funds set aside for cooperation with other national CLARIN projects. Another option to implement the cooperation is to seek funding together in Europe, but, obviously, acquiring such funds will not be easy.

Within the Netherlands, we closely cooperate with DARIAH[39] (Digital Research Infrastructure for the Arts and Humanities). It has resulted, inter alia in a common proposal for the 2011 Dutch roadmap for large scale research facilities, called CLARIAH.[40] There is also close collaboration with the KNAW eHumanities programme[41], and CLARIN-NL is involved in various new project proposals, e.g. Nederlab[42], which intends to set up a range of textual data and tools in the CLARIN infrastructure for carrying out diachronic research on the Dutch language, literature, and culture.

5. Recommendations

In this section we summarize the major recommendations derived from our experiences in the CLARIN-NL project, and we refer to the section where the recommendation was brought up:

- Organize national CLARIN activities as a mix between a programme and a project (cf. section 3.)

  - Organize the implementation of the technical infrastructure as a small set of tightly structured longer-term projects with selected expert groups as participants (cf. section 4.1.)

  - Organize a large part of the resource curation and demonstrator projects as a programme with calls for (relative small and short-term) projects that are open to all humanities researchers (cf. section 4.2.).

- Ensure involvement of the targeted users of the CLARIN infrastructure, e.g. by requiring that they be a partner or even coordinator in a resource curation or demonstrator project. (cf. section 4.2.)

- Organize closed calls (on invitation only, and no competition) for under-represented disciplines and subdisciplines to ensure coverage of all humanities researchers (cf. section 4.2.)

- Create expertise and achieve synergy by setting up dedicated (virtual) centres or teams for specific topics that are crucial for realizing the CLARIN infrastructure, e.g. resource curation, metadata, semantic interoperability:

  - Make sure that there is systematic coordination for semantic interoperability issues, e.g ISOCAT. (cf. section 4.5.).

---

[34]See `http://www.clarin.nl/calendar/` for an overview.

[35]`http://trac.clarin.nl/trac`

[36]See `http://www.clarin.nl/node/185`

[37]`http://www.ccl.kuleuven.be/CLARIN/`

[38]`http://www.clarin-d.de/`

[39]`http://www.dariah.eu/`

[40]`http://www.clariah.nl`

[41]`http://ehumanities.nl`

[42]`http://www.diachronie.nl/`

- Have an expert or expert team create metadata components and profiles and test them against actually existing local data as one of the first activities in the project. Do the same for metadata for software. (cf. section 4.1.)
- Set up a team or centre dedicated to the curation of resources (cf. section 4.3.).

- Reserve funds in national CLARIN projects for cooperation with other national CLARIN projects, on topics to be determined together (cf. section 4.9.).

- Increase sustainability by involving libraries, data archives, and other existing national data centres in the CLARIN infrastructure (cf. section 4.6.)

- Initiate an integration project that has to integrate all components of the infrastructure, and do so sufficiently early, so that problems encountered can still be addressed. (cf. section 4.1.)

6. Conclusions

In this paper we have described some recent developments in the CLARIN-NL project with the goal of sharing information on and experiences in this project with the community outside of the Netherlands. We have provided several recommendations for related projects, e.g. other national CLARIN projects. We hope that other project may benefit from the experiences gained in the CLARIN-NL project.

7. Acknowledgements

8. References

D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry- and component-based metadata framework. In 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. European Language Resources Association (ELRA).

M. Kemps-Snijders, 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`.

Jan Odijk. 2010. The CLARIN-NL project. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 48–53, Valletta, Malta, May. European Language Resources Association (ELRA).

M. van Gompel, M. Reynaert, and A. van den Bosch. 2011. CLAM: Computational linguistics application mediator. Presented at CLIN21, Gent.

M. van Gompel. 2012. FoLiA: Format for linguistic annotation. Presented at CLIN22, Tilburg.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.