CLARIN Centers for a Sustainable Infrastructure

Daan Broeder, Jan Odijk
Max Planck Institute for Psycholinguistics, University of Utrecht
E-mail: daan.broeder@mpi.nl, j.odijk@uu.nl

Keywords: clarin centres, governance, sustainability

1. Introduction

This paper is about the role of centres in the CLARIN research infrastructure.

CLARIN [1] is a research infrastructure initiative that aims at providing a single domain of Language Resources (LR) and Language Technology (LT) to researchers from the Social Sciences and Humanities (SSH) disciplines. CLARIN is on the ESFRI roadmap and was awarded an ERIC in 2012. The basis of the CLARIN infrastructure was already laid in the so-called preparatory phase, with funding from the EC in the CLARIN EU project, that ran from 2008 to 2011. In this project phase, the CLARIN infrastructure organisational and functional architectures was designed and was for a large part to be built on a foundation of CLARIN centres that (1) commit to responsibility for delivering their usual infrastructure services in a CLARIN compatible manner (CLARIN B-type centres and services) and (2) for a subset of centres, commit to delivering general infrastructure services beneficial for the whole CLARIN community (CLARIN A-type centres and services). Usually CLARIN centres are national (research) institutes or University departments involved with linguistic research that also have a role in providing language type data for the research community.

Although the CLARIN building phase that followed the EU preparatory phase was directed by the different national CLARIN projects, the original idea of the CLARIN infrastructure being upheld by a backbone of CLARIN centres has persisted.

Although this centre-oriented approach is not unique for research infrastructures, even in the SSH as in CESSDA [2], other approaches are also widespread. For instance the virtual competence centre approach from DARIAH [3] that pools services and efforts from different centres in sub-units, and single- or few-centre approaches, where the discipline is dominated by a only a few centres as in SHARE [4] and ELIXIR [5].

In this paper we want to present some views about the suitability and success of the centre-centric CLARIN approach, mostly based on the situation in the Dutch CLARIN project, but also informed by the situation in the other national projects. Our views culminate in four recommendations that will in our opinion lead to a better sustainable and persistent infrastructure:

1. CLARIN conformant services should be the basis of also the internal institute's workflow.
2. Include bigger, (multi-disciplinary) research institutes and data service providers
3. Use services that are easy to reallocate.
4. Reserve resources for outsourcing services.

2. CLARIN centre services

The types of services that are currently required from CLARIN centres are mainly of a technical nature, including:

- Archiving and PID management
- Metadata domain
- Services domain
- Common AAI

Each CLARIN-centre has its own focus and provides data and services in accordance with its focus, which makes each CLARIN centre unique. However, the CLARIN-centres have in common that they provide their customary services in a CLARIN-compatible way.

In addition to providing technical services, CLARIN centres are expected to let their staff participate in the CLARIN committees and task forces (and their national equivalents). Although it is not a formal requirement, it is nevertheless an essential assumption that the development of the CLARIN infrastructure should be a communal effort.

A special procedure has been set up to assess whether a CLARIN centre provides its services in a CLARIN-compatible manner. Centres that meet the requirements can become CLARIN-certified centres. In addition to that, it is required that centres are certified in accordance with the Data Seal of Approval (DSA) [6]. This certification requires the centres to make the relevant internal workflow explicit in documentation, and to work according to these documented procedures.

Currently there is a (ERIC-) coordinated move to balance the availability of CLARIN-A type services from CLARIN centres with running such services also at central computing centres. This makes such services more independent from individual institute dynamics but also contributes to professionalising the service deployment context and increasing the service availability. All this comes at a cost and can deeply influence the original community oriented service and resource-provisioning model, making CLARIN more and more dependent on central funding. An alternative can be to attempt to involve big data and computing centres already at the community (CLARIN) level and making

them complicit in the planning and providing a slot for them in the CLARIN centre taxonomy (4)

3. CLARIN centre certification

One of the first priorities of the central CLARIN coordination body has been the formation of two committees necessary for an assessment of the different centres participating in CLARIN. These are the CLARIN centre committee (CCA), necessary to have centres participate in decisions about centre certification requirements, and the CLARIN centre assessment committee (CAC) responsible for the certification itself.

Until now, a grand total of 14 centres have been certified, and although in general the certification procedure seems to work smoothly, some points should be made.

Currently the certification is directed mainly to a single type of centre, the so-called B-centres. Though this type seems to be the prevalent type, it does not fit well for centres and organisations that have a broad mission and also have obligations to projects and users not directly related to CLARIN such as libraries and national-archives. Becoming a certified CLARIN-centre requires more effort from them than for organisations focused on LR & LT, because they have to fit in the CLARIN-compatible way of working with their overall way of working.

Although different classifications are possible, such types are currently non-certifiable and such centres are not felt to be fully part of CLARIN.

A justified question is how CLARIN can keep quality standards with respect to B centre services if it needs to accommodate and involve such different types of centres, either for purposes of long-term organisational sustainability or for political reasons. Especially including broad mission type of organisations can be important for the embedding of CLARIN in the national landscapes.

This point transcends of course the certification on the basis of pure services, and pertains more to the challenge of how to create a harmonious collaborating group of disparate members, and keeping all motivated to deliver their part.

4. Centre Taxonomy

Already in the preparatory phase one was aware of the necessity to accommodate different classes of CLARIN centers as:

- A centres that provide also services for the whole CLARIN community e.g. CLARIN metadata schema registry
- B centres that offer access to data and tools specifically via CLARIN type metadata
- K centres offering expertise and guidance

Currently in the official ERIC CLARIN center taxonomy some changes have been made such as the introduction of "External" E-type centers that offer general services used by the CLARIN community, but are not part of CLARIN itself.

In the national CLARIN projects also local CLARIN centre policies can be implemented. In the Dutch situation we have tried to provide for broad multi-disciplinary centres mentioned in (3) by introducing the CLARIN *Data Provider* centre type, which applies to organisations that, by their mission statement, make data and services available, independently of CLARIN. Such centres need to fulfil only a limited set of requirements, and only for that part of their collections that are relevant to CLARIN. Concrete examples of such centres are libraries (including university libraries), cultural heritage organisations, and archiving institutions. In the Netherlands, the National Library (KB), the Digital Library for Dutch Literature (DBNL), Utrecht University Library Netherlands, and the Netherlands Institute for Sound and Vision became such CLARIN Data Providers.

However the problem remains that such centres can only spend a limited part of their energy at being part of the CLARIN infrastructure i.e. CLARIN is often just one of the many projects they participate in, which only increases questions about any commitment after a period of subsidized participation ends. Therefore some type of certification and commitment should perhaps be considered before subsidizing such collaborations.

5. National specificities

Depending on the national situation, there are large differences in the work load between CLARIN centres with respect to providing the basic infrastructure functionality. For instance, in Finland the CLARIN functionality is completely provided by the University of Helsinki / CSC combination, which is the major Finish academic data & computing centre, a role which they also fullfil for other discipline's research infrastructures and projects. Such a situation of course allows for a maximum of synergy, but in such a context any proposed specialization for LR & LT will need to be argued strongly when compared to general solutions.

In countries with a smaller national CLARIN effort there is a general tendency to support one single national CLARIN B centre only, which is the minimal requirement for participation in the CLARIN consortium. This is an efficient strategy in view of the limited resources, but sometimes also a consequence for smaller[1] countries with few suitable candidate CLARIN centre organizations. The bigger national projects as CLARIN-NL and CLARIN-D that had an early start, support several centres each, with sometimes overlapping interests. In the future we might expect some consolidation there too.

In some countries, the national libraries, which can also

[1] Smaller in the sense of research budget, not necessarily in size

be seen as 'multi-disciplinary' organisations mentioned in (4), play an important role in the CLARIN organisation. From the point of view of sustainability, it is good to include such organizations, which, by their very nature and position in the national landscape, are pretty stable. On the other hand, they are also often perceived to be not technologicaly advanced, not sufficiently e-minded or too much oriented versus publications. This situation is changing fast: libraries are currently, by necessity, reinventing their mission and also entering the data-management domain (cf. data in `enhanced publications', and the increasing requirement to publish the data an journal article is based on). Concomitantly, they are also refreshing their staff and expertise. However this still differs greatly from country to country. It would be highly advantageous if the CLARIN message would be taken up by those advanced libraries and brought into the EU library organizations.

6. Sustainability

Service sustainability was initially expected to be made possible by an initial subsidized phase that would allow centres to make such services CLARIN-compatible and anchor them in the organization's workflow. It was expected that after that the overhead of also delivering those services outside the centre would be of an 'acceptable' level. This assumption is now being challenged, but still underlies the CLARIN basic strategy.

In the Dutch project, CLARIN centres were required to sign a statement that they would indeed anchor such CLARIN compatibility in their own workflow. Similar explicit or implicit expectations are present in other national projects.

We initially expected CLARIN centres in the form of (multi-disciplinary) research institutes to be more persistent than the university research departments and certainly more so than temporary projects. However, in the Dutch project we have now seen also organizational, research political and funding dynamics that make it clear that the stable research institute assumption is at least challenged. This does not only impact the actual CLARIN services but also the network of expertise that was built up in the last years. Therefore we think there should be a twofold strategy:

First make sure that any CLARIN community-wide services are not bound to specific centres and can be reallocated without many problems. Reallocation can take place to other research institute type centres, but also to computing centres or even commercial providers. Suitable resources will have to be made available in the latter case. Second, try including larger centres selecting them on the basis of sustainability (libraries), research organizational embedding (national data centres) etc., Even if such organizations are not purely linguistic oriented, we should do so for the sake of long term stability and attempt to accommodate and embed them in the best possible way.

A separate matter in this all is whether the incorporation of such institutes should influence scope and ambitions of CLARIN itself. On this subject we claim that the focus on Language Resources and Language Technology played an important role in the success of CLARIN. Dilution of the scope is probably not beneficial although for the larger multi-disciplinary centres, an inspirational and guiding aspect of CLARIN as a good SSH example can be most helpful.

All in all we come to the following recommendations:

1. CLARIN conformant services should be used also as a part of the service hosting institute's research workflow to increase the hosting institute's commitment and improve the service persistency.
2. Include and put more emphasis on participation of bigger, (multi-disciplinary) research institutes and data service providers (including libraries).
3. Ensure that services are easily reallocated to diminish the dependency on specific centres.
4. Reserve resources for outsourcing services.

7. References

[1] Váradi, Tamás and Krauwer, Steven and Wittenburg, Peter and Wynne, Martin and Koskenniemi, Kimmo. CLARIN: Common Language Resources and Technology Infrastructure. . LREC. European Language Resources Association, Year 2008.

[2] http://www.cessda.net/

[3] http://www.dariah.eu/

[4] http://www.share-project.org/

[5] http://www.elixir-europe.org/

[6] http://www.datasealofapproval.org/en/