Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure

P. Wittenburg*, N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, L. Lemnitzer, B. Maegaard, M. Piasecki, J.M. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R.v. Veenendaal, T. Váradi, M. Wynne
*MPI for Psycholinguistics
E-mail: peter.wittenburg@mpi.nl

Abstract

Currently, research infrastructures are being designed and established in many disciplines, all partly to address the problem that they all suffer from an enormous fragmentation of their resources and tools. In the domain of language resources and tools the CLARIN initiative has been funded since 2008 to overcome many of the integration and interoperability hurdles. CLARIN can build on knowledge and work from many projects that have been carried out over many years, and will build stable and robust services for use by researchers. Service centres will play a key role, and must provide persistent and highly available services that adhere to criteria as they have been established by CLARIN. In the last year of the preparatory phase of CLARIN these centres are currently developing four use cases that can demonstrate how the various pillars CLARIN has been working on can be integrated. All four use cases fulfil the criteria of being cross-national.

1. Background

Almost all theoretical aspects that have to do with integrating language resources and making them interoperable have been studied in a number of projects already, ranging from projects dealing with syntactic harmonization to those which have tackled semantic mapping. We can refer here to numerous European projects such as, for example, EAGLES [1], MILE [2], ISLE [3], Schemas [4] and Ontolex [5], not to mention the number of national programs. Academics often tend to believe that they have already solved the problem when they have understood the theoretical underpinnings and created a prototype. One of the most often heard phrases is currently that "people have solved this already". Using such an argument they fail to appreciate that this does not mean that there are services available that can be used by non-technical researchers in their daily work. Deprived of a large group of helpers that can adapt the resources and tools to accept certain formats and tag sets, these researchers still face the problem that there are only fragmented services yet that are offered by institutions that cannot guarantee persistence.

We still live in a "download first" scenario where researchers need to spend a lot of time downloading resources and tools, installing them, adapting formats etc., which is the opposite of the e-research scenario that we foresee for the future. There are a number of reasons for this, such as no trust in the availability of stable services, no trust that web services will be secure and no chance to solve the integration and interoperability problems with the help of existing web services. Moreover, the existing web services often have an experimental character and have not been designed for intensive use by hundreds of users and large volumes of data. The huge challenge for research infrastructures such as CLARIN [6] is therefore not to come up with new theoretical insights about interoperability solutions, but rather to turn existing knowledge into trusted, stable and robust services that researchers would finally accept and rely on. In this respect we should not forget that today's research scene is very competitive, since only those who publish unique results of high quality will receive grants. Having said this it also should be clear that it is an illusion to think that a change in researchers' attitudes will occur within a short time frame. We will probably need to carry out much training and education effort in order for this to happen.

Establishing a research infrastructure that offers services of the kind indicated above requires advanced and highly specialized technological expertise not commonly available in ordinary research centres. However, as a result of years of discussions and debates about integration and interoperability at forums such as, for example, LREC [7] and ISO TC37 [8] the domain is arguably mature enough to establish the pillars of an infrastructure overcoming this fragmentation. The question that needs to be addressed, then, is who will be the players that can guarantee availability (in terms of permanent working and an appropriate level of processing efficiency), that take care of persistence of resources and services and that can promote step by step the use of standards to ensure interoperability. Obviously we need centres that offer services in the described manner and we need to motivate researchers to rely on these services and to allow these services to be integrated in their workflows. But, at the same time it is clear that these centres need to be in close communication with the relevant research institutions in order to keep resources and services informed by ongoing relevant research.

2. Requirements

CLARIN will build a backbone network of centres that will offer the services based on formal statements about their quality and duration. The LRT community can certainly learn from the discussions around the Large Hadron Collider [9] and the DEISA project [10]. Centres have been identified that will be able to guarantee certain services based on service level agreements augmented with an extensive support concept. Any infrastructure that wants to be taken seriously by researchers will need to come to a similar framework. In addition centres that want to participate in the backbone network of the CLARIN infrastructure must fulfill a number of different criteria [11] out of which we mention only the most important ones in this paper:

1) a proper repository system is required that will resolve unique and persistent identifiers into the intended resource or resource fragment and that will be checked regularly according to a quality assessment procedure as indicated by the Data Seal of Approval [12] method;
2) adherence to all CLARIN specifications about standards and protocols, for example, with respect to providing high-quality metadata and resource formats [13];
3) participation in the national identity federations (where available, otherwise the centre should cooperate with initiatives in building such federations) and in the CLARIN service provider federation [14], all based on widely accepted protocols and trust agreements;
4) explicitness with respect to IPR, license and ethical issues, and specification of the "business model" so that depositors and users know what they can expect;
5) opportunities for researchers to integrate the offered resources into their workflows without the current technical restrictions defined by the specific search engines.

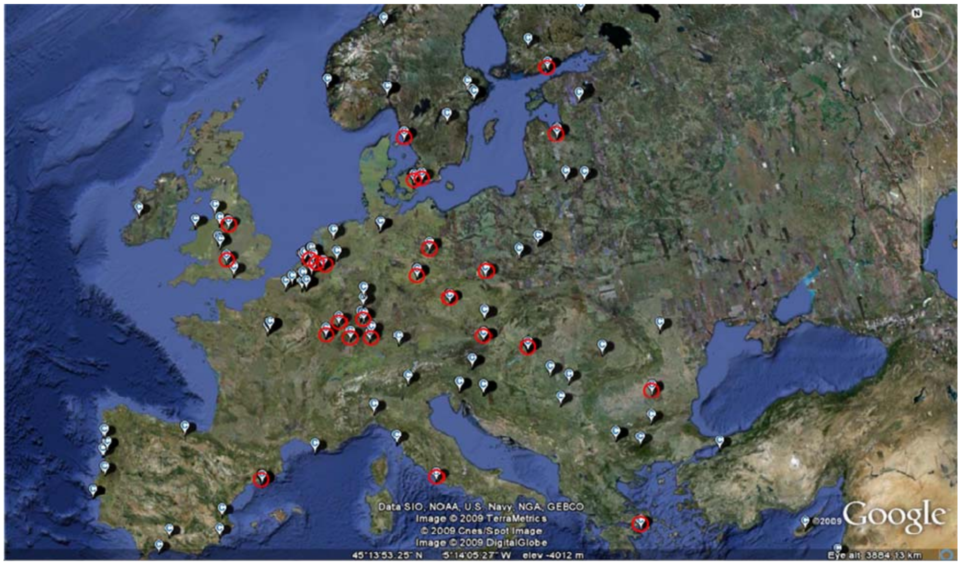

Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2009 TerraMetrics
© 2009 Cnes/Spot Image
Image © 2009 DigitalGlobe
45°13'53.25" N 5°14'05.27" W elev -4012 m
©2009
Google
Eye alt 3884.13 km

Figure 1 indicates those institutions that have indicated an interest to act as CLARIN centres.

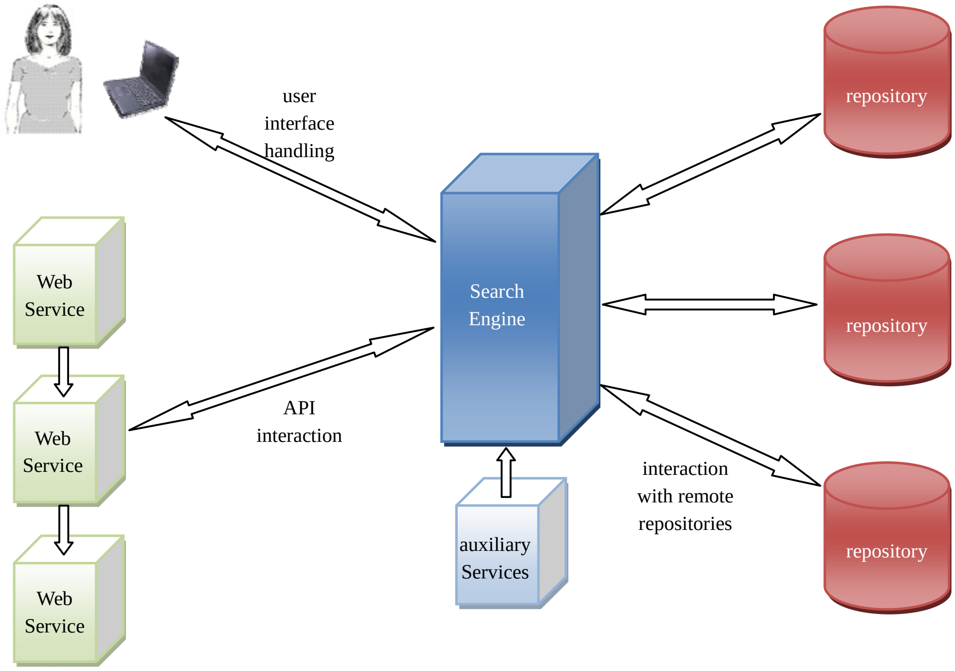
In particular the adherence to standards for resource formats and linguistic encoding is of greatest importance. Therefore CLARIN has agreed on a variety of existing and emerging standards such as ISOcat [15] as a reference vocabulary for linguistic encoding, LMF [16] as a common format for lexical resources, corpus standards such as LAF [17], SynAF [18] and SemAF [19] and best-practices in the area of multimedia/multimodal annotations [20]. It is obvious that many of these standards are new, but a research infrastructure such as CLARIN is committed to testing and helping to develop these standards to maturity. As pivot formats they may play an important role in the emerging integrated scenario.

Most of the above-mentioned criteria apply to typical centres offering resources. However, similar requirements can be established for centres offering language technology services such as those based on, for example, Natural Language Processing or Automatic Speech Recognition components. An important difference is that service centres will have to face the potentially growing interest of the future users. This issue is even more important as users from across the Humanities and Social Sciences disciplines often do not have access to computing resources and do not have enough technical skills to configure and use language tools on their own computers. Thus, service centres should guarantee an appropriate level of computing power to serve such users. For some tasks, e.g. shallow statistical semantic analyses of language in large corpora the needs for processing power can be substantial even in the case of one user performing an analysis, not to mention many users working simultaneously. This issue brings in organisational and technical aspects similar to those well known in the domain of computing centres in e.g. chemistry and physics, but it is still very seldomly discussed in the context of language technology infrastructure. Recently, a large European initiative called PARADE [21] was presented that will tackle these issues by creating a horizontal data services e-Infrastructure.

Currently, more than 20 institutes from the majority of European countries (see figure 1) have accepted these requirements, indicated their interest to participate and therefore to adjust their local setup including the acceptance of following agreed standards where possible. A first assessment at the beginning of 2009 revealed that most of the candidates need to take considerable measures to meet the criteria. A second assessment in October 2009 showed that almost all are currently working intensively on developing their services. The speed at which these adaptations take place varies, depending on the amount of funding support, but in all cases there is a clear interest in living up to the criteria.

3. Demo Case

A network of centres based on federation principles, as described above, can only show its added value by means of integrated demonstration cases. Therefore, CLARIN partners including a number of centre candidates have been defining a few use cases that can show how many of the pillars of the intended infrastructure will work together. The basic design of these cases is indicated in figure 2. It will include access to resources stored at distributed repositories via a cross-search function, and, where possible, include the execution of web services for example on the result set. Four demo cases have been identified:



- A multimedia and multimodal use case where cross-repository metadata and content queries on interesting collections should be possible. The major aspect in this use case is to improve the metadata information and to map between the various types of encodings.
- A second use case will execute typical NLP operations on a corpus (C4) stored in centres from 3 different countries. The NLP operations are made available by a variety of centres as web-services, and the WebLicht framework will allow users to configure the chain of operations which will then be executed at different centres.
- A third use case will exploit newspaper articles from various countries stored at different centres that focus on the same topics and extract an Identity Index by applying typical NLP operations up to Named Entity Recognition. Also in this case web services provided by different groups will be involved.
- A fourth case will focus on fairy tales which are stored at different centres and which will also be analyzed with the help of NLP operations implemented as web services.

In all cases web applications will offer a simple interface to allow users to initiate the corresponding operations.

Since CLARIN metadata is already being harvested centrally, as can be seen from the Virtual Language Observatory [24], the search of resources by metadata can be initiated from a single portal resulting in a selection of resources which are however stored at the various, distributed repositories accessible in CLARIN. In the first use case resources will constitute a transient virtual collection in which the content can be searched. These content searches will either map queries expressed by a user to the search engines offered by the various repositories, or will use a generic search web application. In the latter case queries will be distributed via web service interaction and standardized protocols to the local programming interfaces and receive the hits via the same mechanisms. These hits are actionable in so far as locally available web applications or, again, remote web services can be used to visualize the data. In the same way, other web services can be used to process the temporary collection or the hits generated, including natural language processing chains as those offered by UPF [25], WebLicht [26], GATE [27], for example.

These demo cases will bring together a number of the pillars CLARIN is currently working on. In addition, it will be a first step towards generalizing search functionality and including more web services from different teams all over Europe. The demo cases will be realized in 2010. However, in its first phase it will not yet include advanced features such as structured content search and search for complex time/sequence patterns for example.

4. Trust Domain

In the mean time an initial trust federation has been established that currently covers three countries: Finland, Netherlands and Germany. CLARIN service providers from these countries have signed a contract specifying the terms under which they accept user attributes from identity providers. This service provider federation also signed contracts with the corresponding national identity federations in the participating countries. This trust domain based on legal documents includes now more than 1 million researchers and students as potential users that can access the provided resources by using a single sign-on mechanism once authenticated by the research institution they are employed at. This feature can now being used to access resources from the participating centres in a simple way. Once tested it is the intention to extend this initial federation quickly to other centres that have integrated Shibboleth (or SimpleSAMLPHP) components in their repository and to countries with national identity federations. The intention is to use this mechanism also in those use cases where access is being done to resources from different centres from a web application.

In the cases of web services we will not yet be able to implement this single sign-on mechanism, since currently there is no technology available to pass user credentials gained by a Shibboleth based distributed authentication step over to web services. The Grid domain is relying on X509 based user certificates, however, such methods cannot be introduced in the humanities. It is the intention to collaborate with the Grid experts to work on a special implementation of so called short-lived certificates to implement single sign-on for web-services. But this will not be finished in the preparation period of CLARIN.

5. Summary

We have presented the need to establish a backbone network of centres that will adhere to a number of requirements. In particular, warrantee for high availability and persistency of the offered resources and services and adherence to agreed standards will help researchers to avoid the current "download-first" paradigm that does not exploit the full potential of the web. Four concrete use cases have been designed and selected that integrate resources and web-services from centres from different countries to demonstrate the potential of eScience methods. In the preparatory phase not all wishes can be realized of course. For example for web services there is no technology available to implement single sign-on principles if one would not step over to X509 based user certification which is not acceptable in the humanities.

The chances are high that CLARIN funding will continue so that the construction phase can be started. Prototypical solutions that have already been worked out will then be integrated and extended to overcome the many hurdles which we are faced with.

6. References

[1] http://www.ilc.cnr.it/EAGLES96/guide/node17.html
[2] http://www.ilc.cnr.it/EAGLES/isle/program.htm
[3] http://www.mpi.nl/ISLE/
[4] http://www.schemas-forum.org/
[5] http://www.ontotext.com/OntoLex/
[6] http://www.clarin.eu
[7] http://www.lrec-conf.org/
[8] http://www.tc37sc4.org/
[9] http://lhc.web.cern.ch/lhc/
[10] http://www.deisa.eu/
[11] http://www.clarin.eu/files/wg2-1-centers-doc-v8.pdf
[12] http://www.datasealofapproval.org/
[13] http://www.clarin.eu/system/files/private/Standardisation%20action%20plan-v8.pdf
[14] http://www.clarin.eu/files/wg2-2-federation-doc-v6.pdf
[15] http://www.isocat.org/
[16] http://www.lexicalmarkupframework.org/
[17] http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf
[18] http://www.tc37sc4.org/new_doc/ISO_TC37_4_N244_SynAF_WD_draft.pdf
[19] http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf
[20] Thomas Schmidt: AG/EAF/Exmeralda etc.. NEERI 09 Conference 2009, Helsinki
[21] http://www.csc.fi/english/pages/neeri09/programme/index_html/?searchterm=identity
[22] http://www.edugain.org/
[23] http://www.mpi.nl/DAM-LR/
[24] http://www.clarin.eu/vlw/observatory.php
[25] http://www.clarin.eu/node/1471
[26] http://www.sfs.uni-tuebingen.de/dspin/weblicht.shtml
[27] http://gate.ac.uk/