Building a federation of Language Resource Repositories; the DAM-LR project and its continuation within CLARIN

Daan Broeder, David Nathan, Sven Strömqvist, Remco van Veenendaal
MPI for Psycholinguistics, SOAS University of London, Lund University, Institute for Dutch Lexicology
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: daan.broeder@mpi.nl, djn@soas.ac.uk, sven.stromqvist@ling.lu.se, veenendaal@inl.nl

Abstract

The DAM-LR project aims at virtually integrating various European language resource archives that allow users to navigate and operate in a single unified domain of language resources. This type of integration introduces Grid technology to the humanities disciplines and forms a federation of archives. The complete architecture is designed based on a few well-known components .This is considered the basis for building a research infrastructure for Language Resources as is planned within the CLARIN project. The DAM-LR project was purposefully started with only a small number of participants for flexibility and to avoid complex contract negotiations with respect to legal issues. Now that we have gained insights into the basic technology issues and organizational issues, it is foreseen that the federation will be expanded considerably within the CLARIN project that will also address the associated legal issues.

1. Introduction

This paper reports on the results of the DAM-LR project (Distributed Access Management for Language Resources) which was introduced at LREC 2006 [1, 2] and will also focus on the plans for continuing the DAM-LR within the CLARIN (Common Language Resources and Technologies Infrastructure) project [3]. The DAM-LR project aimed at investigating the needs and laying the basis for establishing a federation of language archives that can offer their users a shared virtual space of resource location facilities and resource access.

CLARIN is committed to establishing an integrated and interoperable research infrastructure and technology for language resources and its technology. It aims at removing the current fragmentation, to offer a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities. So it is far more ambitious than DAM-LR but will be able to build on the experiences of DAM-LR.

An overview of the used technologies used and the created infrastructure will be given in the context of the learning processes that informed them. Also, included is an inventory of non-technological issues that were encountered and that will need to be addressed also within CLARIN such as establishing the level of support an organization needs to maintain this kind of infrastructure and the need to establish contracts with clear responsibilities.

2. Federation Participants

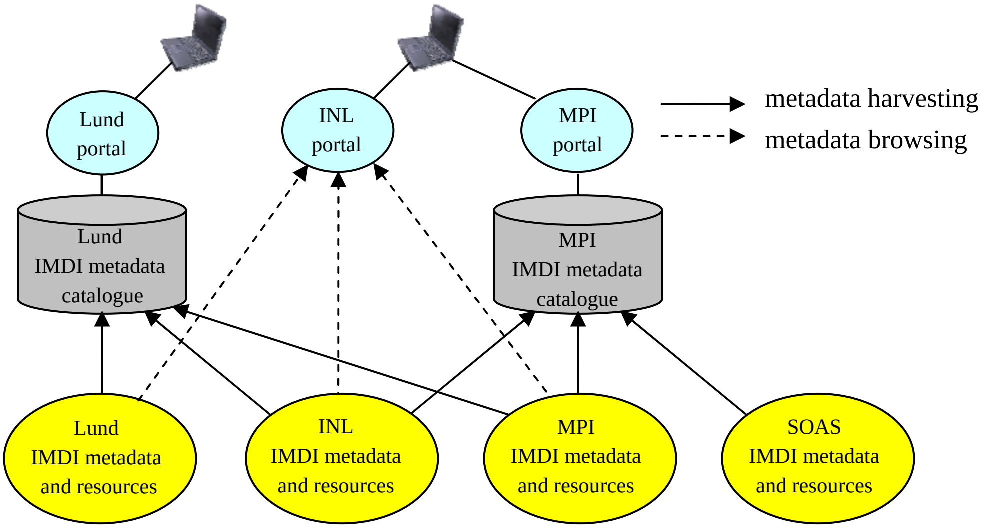
The group of DAM-LR participants: (1) Lund University, (2) SOAS University of London, (3) Institute for Dutch Lexicography (INL) and the (4) Max-Plank Institute for Psycholinguistics (MPI) - was purposefully limited to allow for easy communication and the need for rapid adjustment of the plans when needed. This is an important issue when exploring the uses of unfamiliar technologies. However, it is important to state that this work took also place within the broader context of the DELAMAN network of archives [4] that focuses on the preservation of endangered languages and music. and the Regional Archives initiative [5] which was initiated by the MPI for Psycholinguistics (supported by the Volkswagen Foundation within the realm of the DOBES project and funded by the Max-Planck Gesellschaft to establish regional archives of language material in the countries of origin). Both these contexts guided our designs. Although the initiative is primarily directed towards endangered languages and cultures the regional archives can hold a much broader type of digital material because the underlying repository structure is not limited to language resources. This may be welcome in view of the lack of archiving facilities in many of these places.

3. Federation Objectives

The work of DAM-LR focused on three pillars:

1) Building a shared searchable and browsable metadata domain.
2) Creating an identity federation that allows each partner archive’s users to access the other archive’s resources while leaving authentication to the user’s home archive
3) A unified space of resources identified by unique and persistent resource identifiers. This allows persistent referencing and citation, by treating different instances offered at various sites as the same resource.

All mentioned integration services need to be based on signed certificates accepted by instances such as TERENA [6] to prevent misuse.



Lund portal
INL portal
MPI portal
metadata harvesting
metadata browsing
Lund IMDI metadata catalogue
MPI IMDI metadata catalogue
Lund IMDI metadata and resources
INL IMDI metadata and resources
MPI IMDI metadata and resources
SOAS IMDI metadata and resources

Figure 1 Shared metadata domain. Metadata is harvested from all participants in searchable & browsable catalogue systems by those sites that support it (Lund & MPI). INL supports a browsing only catalogue of the available metadata. SOAS offers metadata for harvesting, but relies on its own catalogue system (not shown).

Shared Metadata Domain

We were successful in creating a single metadata catalogue by harvesting each others metadata records using IMDI metadata [7, 8] and a simple HTTP crawling method. Although we support the OAI-PMH Protocol for Metadata Harvesting [9] as a metadata dissemination protocol it was not necessary to also harvest metadata reusing this protocol since all parties were able to provide the metadata as an IMDI corpus tree. We were helped by the fact that most partners already used IMDI metadata (or an older version of it) for their internal information systems; making mapping to the agreed exchange set (IMDI) was relatively easy. When up-scaling to larger federations this will not be that easy. Within CLARIN there will be an expansion of the type of resources and using IMDI as the only metadata exchange set will no longer be possible. To address the different needs and terminologies, CLARIN plans to adopt a flexible component approach to metadata creation, metadata mapping and exchange. In that context the metadata terminology mapping work (metadata crosswalk) will be supported by references to an ISO T37/SC 4 DATCAT [10] data category registry service and the exchange of metadata is expected to use the OAI-PMH protocol.

Identity Federation

The DAM-LR identity federation is based on Shibboleth [11], a collection of middleware components that allow users to authenticate with their home organization when they want to access resources stored with other federation partners. The organization storing the resource keeps complete control over the access policy. It bases authorization decisions on a number of user attributes that are provided by the user’s home organization. Determining a set of user attributes appropriate for the

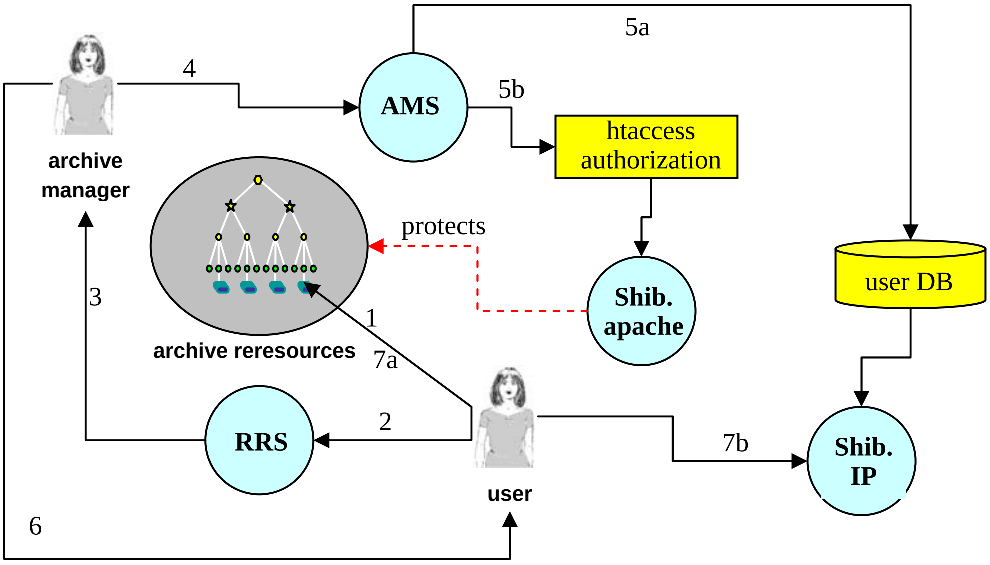


Figure 2 Access management and Shibboleth

domain of language resources is an important part of this work as well as learning how to configure and integrating the maintenance of authentication and authorization mechanisms into the normal operations of the archive.

In Figure 2. the integration of Shibboleth with the access management procedures and components of the MPI reference solution is shown. An unknown user attempts to access via a shibbolized Apache web server a protected resource (1). When she has no access she is referred (2) to the Resource Request System (RRS) that sends a message to the archive manager (3) who then decides about access. If she assents she uses (4) the Access Management System (AMS) to add a user record to the user data base (5a) which adds an authorization record to the htaccess file (5b). The user is then warned (6) that she can access the resource (7a) after she authenticates with the Shibboleth Identity Provider (7b) with her identity.

These technical issues are not trivial and we expect that not every archive will support them. In the context of building "regional archives" where professional support people are often more rare, these functions should be optional, to be supported by those who have the capability. In the future it is possible better configuration tools will offer a solution. Within CLARIN we plan for a special support team to help and train local staff. Although Shibboleth is an excellent solution for accessing resources via a standard web browser, many usage scenarios in the language resource domain require access from applications or (web) portals. A possible solution may be the integrating of Shibboleth with the issuing of short lived credentials (SLCS CA), something that is done by the Swiss Science Network [12]. We hope to realize a similar solution in cooperation with the Dutch e-science grid project within CLARIN.

Persistent Resource Identifiers

The third pillar - using persistent resource identifiers (PIDs) that are independent of the actual resource location - was also successfully completed. As an implementation of this concept, DAM-LR chose for the Handle System (HS) [13] that is already well established in the digital library world. Configuring and maintaining this system adds to the archives workload but less so than the Shibboleth middleware. No special domain specific issues were actually worked on within DAM-LR, but working with the PID concept and the Handle System did make us conscious of the need to have stable references not only for complete resources but also for resource fragments. This would allow, for instance researchers to establish relations between and make commentaries on annotation fragments. We have decided to pursue this work within the context of ISO TC37/TC4 in order to propose a standard for referencing of web-based resources and fragments.

Some work could not be completed although preparatory work was done. It is particularly important within the context of establishing "regional archives", to be able to synchronize datasets between archives. Not all archives behave good data back-up facilities or high network bandwidth, so it becomes important that data can be synchronized with holdings at archives that have more secure storage. Also alternative resource locations with better network bandwidth could ease the dissemination burden on less well equipped archives. Within the "regional archives" initiative as well as in CLARIN, further developments will create possibilities for synchronizing datasets in a controlled way, offering data originators fine grained control over the exact content that will be copied to a sister archive.

4. Legal issues

Legal issues tend to be underemphasized in academic environments but are increasingly relevant as facilities become more interdependent in fulfilling core operations, as institutions collaborate across national boundaries, as they are more involved in dissemination of information, and as operations come under laws relating to privacy and data protection.

There are in general few concerns about archive users' privacy and data protection because the identity federation mechanism uses Shibboleth for authentication and does not disseminate users' attributes outside their host institution. On the other hand, it may be more difficult to ensure that any contractual obligations with depositors are maintained. Many archive resources are deposited under restrictive conditions, such as limiting access in various ways, so that federation partners acting as an access portal could be exposed to problems caused by the originating site not correctly implementing such access restrictions. In addition, language materials are often recorded spontaneously in informal settings, and potentially have libelous, inflammatory, or even incriminating content making those who deliver such materials vulnerable. Furthermore, under the "Live archives" proposal, future archives may allow users to "add layers of interpretation, annotation and commentary" [1], blurring the boundary between depositors and users and providing more opportunities for contentious scenarios. Such scenarios are compounded in the distributed world of DAM-LR where the unclear notion of "publisher" may be a source of concern (as indeed it already is on the WWW).

One important means of providing protection for archives and depositors is through end user agreement forms, where users register their agreement to conditions before being permitted to download resources. These agreements are, for the user, the point where interacting with the federation gateway gives way to the interfaces and rules imposed by the local archive. In DAM-LR we have drafted an end-user agreement mechanism but have not fully explored its implications.

At the highest level, legal obligations (and legal problems that arise) need to be handled through a governance structure. In DAM-LR, governance is not formal but emerged from participants' shared goals and activities within the project and its funding structure. This is not a very sustainable model, especially in the case of archives with a long term view of operations, since the participants' shared obligations are constrained within the scope of the funded project, rather than serving the broader goals of archiving in the language domain. Governance considerations will be of various types: technical (a managed set of technical rules, regarding matters which underlie fundamental operations, such as metadata and resource identifiers as described above); financial (the financial environment that provides for joint and individual archive participation at various levels); and strategic (for example, the extent that individual archives may set and change their priorities, such as collection strategies). Mechanisms will be needed to be able to handle legal and financial issues that arise, for example, where resilient long term data preservation relies on interdependence between institutions [15], or when there are major transfers of archive holdings to other institutions. To this end, CLARIN is designed with a more explicit governance strategy than DAM-LR, and includes boards focusing on scientific, co-ordination, and management matters.

5. Lessons for CLARIN

Research infrastructures (RI) such as CLARIN are meant to form the basis for future eScience scenarios in which it is foreseen that researchers can seamlessly access and combine resources and services offered by various service and repository centers. Such RIs need to be based on pillars as implemented in DAM-LR. Except from the field of the big libraries, some grid experts and a few centers that participated in establishing national identity federations such HAKA in Finland, we can state that the awareness and knowledge about basic federation technologies and the agreements necessary is very limited in the research world. This is in particular true in the humanities and social sciences although our needs to understand the basis of our societies requires globalized information and service exchange.

DAM-LR helped increasing the awareness in the field and created a few additional experts that know how to this kind of integration work. National grid initiatives still are much more oriented to help natural and life science projects, DAM-LR helped to shift the focus and humanities centers understand that they need to contact their national grid experts for getting help. This is in particular important, since in the humanities the heterogeneity in service and resource types and in the way repositories have been setup is higher, i.e. more help will be required to integrate all centers.

Important for the discipline is the increased understanding that there will be a move towards strong non-burocratic centers with national support that can offer services in a stable way. It cannot be the purpose that all universities establish themselves as service centers in our domain, researchers need to understand to hand over their digital resources and tools to such centers to make them available. These centers will form a "Language Resource and Technology Federation" that can interact with the national identity federations currently emerging, sign agreements and in doing so allowing the national researchers community to get easy access to the services it offers. Important is that this LRT federation is under the control of the researcher community.

As several others found out as well we can summarize that federation technologies in general is not in the state of "off-the-shelf" so that everyone can easily install them. Special expertise is required which is not wide-spread yet. The lack of standardization for example in the area of metadata and PIDs is also hampering fast progress. CLARIN need to address these issues and needs to come up with a strategy to help centers crossing national boundaries and it needs to collaborate with ISO TC37/SC4 to push forward new standards. In the first phase CLARIN with its 100 members needs to address issues such as scalability of all components and redundancy of all types of infrastructure services to be accepted.

6. References

[1] Broeder, D., Offenga, F., Wittenburg, P., van der Kamp, P., Nathan, D., & Strömqvist, S. (2006). Technologies for a federation of language resource archives. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 2291–2294)

[2] Broeder, Nathan, D., van Veenendaal, R., & Strömqvist, S. (2006). A Grid for Language Resource Repositories *Proceedings Second IEEE International Conference on e-Science and Grid Computing (e-Science'06*) p. 135

[3] http://www.mpi.nl/clarin

[4] DELAMAN, http://www.delaman.org/

[5] http://www.mpi.nl/DOBES/regional_archives/

[6] TERENA, http://www.terena.org/

[7] Broeder, D., & Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction, *International Journal of Metadata, Semantics and Ontologies*, *1*, 119–132.

[8] http://www.mpi.nl/imdi

[9] OAI, Open Archives Initiative: http://www.openarchives.org/

[10] Kemps-Snijders, M., Ducret, J., Romary, l., & Wittenburg, P. (2006). An API for Accessing the ISO Data Category Registry. *Proceedings of the 5t International Conference on Language Resources and*

Evaluation (LREC 2006) (pp. 22991–2302)

[11] http://shibboleth.internet2.edu

[12] http://www.switch.ch/grid/slcs

[13] http://www.handle.net

[14] http://www.lat-mpi.eu/papers/flyers/DLRA_Flyer_2006-05-09.pdf/file_view

[15] http://www.clarin.eu/structure