Integrated Services for the Language Resource Domain

Daan Broeder, Peter Wittenburg, Alex Klassmann, Freddie Offenga

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{daan.broeder,alex.klassmann,freddy.offenga,peter.wittenburg}@mpi.nl

Abstract

Integrated services for the Language Resource domain will enable users to operate in a single unified domain of language resources. This type of integration introduces Grid technology to the humanities disciplines and allows the formation of a federation of archives. The DAM-LR project, will establish such a federation, integrating various European language resource archives. The complete architecture is designed based on a few well-known components and some integrated services are already tested and available.

1. Introduction

Creating integrated services and sharing resources between like minded archives for language resources as described by the “Live Archives” document [1] looks like an attractive proposition.

The aim is to benefit the user by creating an environment that allows access to all archives as one single virtual archive. It will benefit the participating archives as well by allowing them to better serve their users, allow pooling resources and development efforts and improving the basis of long term preservation.

The integration and sharing technologies used for such an effort are often referred to as “Grid” technologies [2], and in the world of hard science they are a popular subject for forming cooperative groups of institutes and archives called “federations”. In the humanities especially so in the language resource domain such initiatives are rare. The work described here is largely developed within the DAM-LR [3] project that is one of the few that aims at establishing such a federation in the domain of language resources. While Grid technology solutions in the hard sciences were mainly driven by the typical compute bound tasks, leading to the development of middleware such as the Globus Toolkit [4], the humanities interests are more in-line with Data Grid solutions mainly inspired and coming from the Digital Library community.

In this paper we will not go into the organizational, legal and other non-technical aspects of forming such federation but leave it with mentioning that trust embodied in some kind of organizational form is required to make it all work.

2. Integrated Services for Language Archives

In many cases when we use the words “integrating” and “sharing” we actually are talking about interoperability. Users see a single domain of searchable metadata but the metadata format itself can be implemented differently for different archives. There is, however, a gateway that connects and translates to the agreed format so a single integrated “shared” domain is presented to the users.

Services that can be shared or integrated between language archives that present substantial advantages to the users are:

1) Sharing a single metadata domain for searching and browsing. This allows users to formulate one single query for “interesting” resources and obtain results of all cooperating archives. The required precision for such queries determined by the research questions also requires a domain specific metadata set. For more general queries more general metadata sets, shared by possibly other domains as well, can be used.

2) Sharing a scheme for persistent identifiers for resources. This is an issue when supporting references to resources stored in archives. It is well known that URLs are not the ideal means to do this. Different schemes for supporting persistent identifiers have been developed in the librarians’ domain: Persistent URLs (PURL) [5] and the Handle System (HS) [6]. Sharing the persistent identifier scheme allows archives to easily reference each others resources and exchange resources with embedded references.

3) Secure authentication of archive identity. When sharing resources it is important to be able to establish the partners’ identities. Without this, agreed access policies for instance, can not be guaranteed.

4) Single sign-on domain. Language Resource archives cater for the same user community. It would be very welcome if a single user identity can be established requiring a user to identify him only once when accessing resources from different archives.

5) Shared access policy or authorization. For reasons of efficiency it can be advantageous to copy resources between archives. It is important that the access policies of the originating archive for that resource are maintained. If also a single user identity domain is shared (see the previous point), this authorization information can be specific at the level of access by individual users.

The above enumeration of shared services does not imply that all of these should be actually shared between all the members of a federation. Indeed an opt-out for some difficult to maintain services can be desirable to also allow the participation of partners not able to maintain such a service. This requires an architectural framework where these shared services are as much independent as possible.

This independence is not to be confused with the possible organizational requirements where for instance it may be required to actually support a specific way of authentication, one that is trusted by the partner institutions. Or a service can be essential to the goals of a federation or project such as supporting a metadata infrastructure so the resources will be visible via a central portal.

The choice for a particular technology to implement the shared services is usually a matter of pragmatics. One of the partners can already have an installed base that can relatively easily be extended and used by other federation partners. However, it is always sensible to agree on the definitions of the exchange protocols rather than defining the implementation technologies. This allows individual archives the freedom in choosing the actual implementation while concentrating on the interoperability issue.

3. DAM-LR integrated services

In accordance with principles mentioned above, the DAM-LR project emphasized agreeing about the use of certain protocols for interoperability, leaving the partners free to choose a different implementation where possible. However the Max-Planck Institute for Psycholinguistics (MPI) agreed to further develop its archive management solution as a “reference implementation” demonstrating the integrated DAM-LR functionality. Some additional Grid components like the HS for persistent identifiers, were chosen especially because of an existing robust and dependable implementation and its already existing user base.

Prerequisite for all accepted solutions is that any integration component can only be accepted when it is distributed and redundant so that every archive can also function completely autonomous. In the following we will introduce the key pillars of the DAM-LR architecture that is also summarized in figure1.

3.1. Integrated Metadata Domain

With respect to metadata interoperability the following principles were agreed upon:

1) The IMDI metadata infrastructure [7],[8] will be supported for browsing and searching either by using the actual IMDI metadata format for storing metadata or by creating them on the fly from a local format or database. At least two portals will be made available with full functionality of metadata browsing and searching.

2) The Open Archives Initiative’s (OAI) PMH [9] protocol is supported to allow harvesting metadata also in DC record format allowing interoperability to the outside world at the level of OAI service providers.

How the different partner archives make use of the integrated domain of IMDI metadata is a matter of choice, the “reference implementation” developed at the MPI and adopted by a number of the partners is described in 4.1.

3.2. Persistent Resource Identifiers

The DAM-LR archives will use persistent resource identifiers or URIDs (Unique Resource Identifiers) to enable stable references for their resources. The problems pertaining to the use of URLs are well known. Previous discussions have shown the advantage of using the Handle System over its contender PURL; the other widely used persistent identifier system. The Handle System of the CNRI [10] provides a highly available service for resolving URIDs to actual URLs. The HS is well known in the library community, adopting it will guarantee stable references from non-local resources (stand-off annotations) and also from publications.

The archive at MPI currently has a HS available for resolving references to its resources. The HS is integrated with other archive services in such a way that it is not an essential service but a highly desirable one.

The DAM-LR partners have agreed to host replications of each others handle service revolvers so this will be a distributed highly available service within the DAM-LR federation. Currently, the simplest scheme was chosen where one partner, possibly the MPI, has copies of all other Handle Systems.

3.3. Secure Archive Identification

All confidential communication between DAM-LR servers and services has to be secure. The interaction between peer components such as for instance those involved with user authentication are based on the existence of a domain of trusted servers and services and each component has to make sure that it is provably identified to be the one that it claims to be. As a means of implementing such a trusted domain, the TACAR list [11] of mutually agreed certificates was created, based on the principles of EUGridPMA [12]. In this implementation, national bodies declare that they will accept certificates form each other, with a Public Key Infrastructure [13] used to sign certificates. Every federation member has to apply to their national Certificate Authority to request the status of a Registration Authority, if the appropriate university is not already a Certification or Registration Authority. Once recognized as a Certification or Registration Authority, sites can issue or request certificates that will be accepted within the EUGridPMA domain.

3.4. Distributed User Authentication

Although all the cooperating archives aim at self sufficiency, several share a group of (potential) users that would like to access resources housed at different places without maintaining different user accounts. Therefore, it would be advantageous if the archives should accept each others identification and authentication of users. An accepted solution for this is the Shibboleth system [14].

The Shibboleth concept is primarily aimed at situations where users can be described by attributes such as “member of university class X”. The authentication of the student is left to the student’s home institution and the others grant access to individual resources based on the attributes associated with his identity. However, for individually operating researchers this scheme does not work as every individual needs still to be identifiable at each site when access rights are determined. In spite of this mismatch of required user specificity, Shibboleth brings the advantage of user authentication being performed at the users home institution and transmitting in a secure way only limited and agreed user information over the internet.

Other possibilities have been considered such as the AAA toolkit [15] that emerged from the Grid community discussions as were also solutions based on a shared LDAP [16] domain. Shibboleth, however, looks to become the most widely accepted standard and might even become a requirement imposed by national libraries, government institutions or funding agencies.

Basically, the partners agree that user management should be done by the home site and that privacy sensitive information such as passwords will not be exchanged. Instead a user will be identified by a unique key that will be transmitted together with a limited number of user attributes between the partners. This key will be used in authorization records when associating resource access policies with users.

3.5. Access Authorization

The access authorization is different from user identification and authentication; it links resource access policies with user and/or group identifiers. If we consider the possibility that archives store copies of each others resources we have to make sure that the access policies remain the same irrelevant of the place where the copy of the resource is stored. Therefore, it seems a natural fit that the authorization records are coupled together with the resource's URID record in the HS. The HS allows to add such user defined record to every handle and thanks to the HS high availability, the authorization record will be available even when the "owner" archive is off-line in the same way as its URID will be.

An access manager component has to be developed or integrated that will match the Shibboleth provided identity with the policy stored in HS record, this can perhaps be achieved by extending Shibboleth's default access manager.

As already stated, the authorization records contain access policies mapped to Shibboleth provided and proven user identifiers and maybe some group access policies, however, Shibboleth does not provide archive managers with authorization records where none yet exists. If a user requests access to a resource this request has to be processed such that the unique federation wide user identifier is confirmed and suitable records can be produced if the archive manager approved the request. Such a resource request management system needs to be developed separately from Shibboleth.

4. Additional functions and Specific Implementation Issues

The following functions and applications are not part of any proscribed DAM-LR integrated service. However, they are essential for running a useful and consistent archive.

4.1. Metadata Utilization.

Within DAM-LR different portals will be established that allow utilization of the integrated metadata domain so users can find relevant resources searching all the partner archives simultaneously. The DAM-LR partners are free to develop their own solution for this, but the majority has chosen to adopt the IMDI infrastructure that allows the

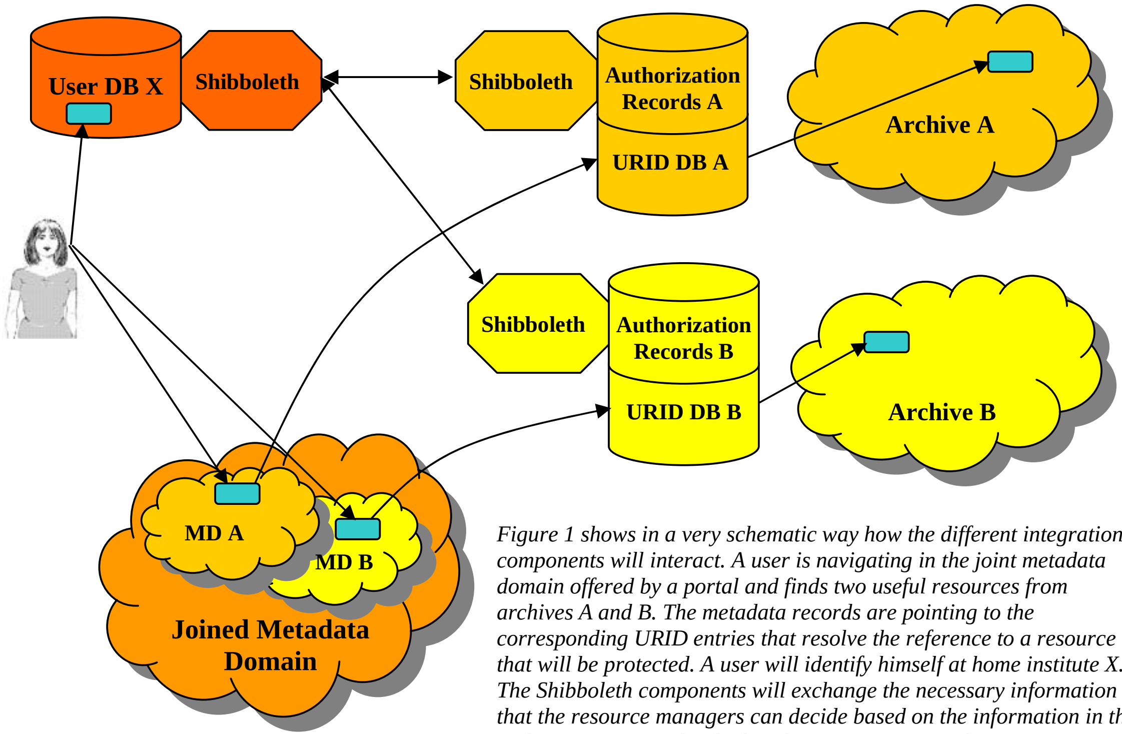

Figure 1 shows in a very schematic way how the different integration components will interact. A user is navigating in the joint metadata domain offered by a portal and finds two useful resources from archives A and B. The metadata records are pointing to the corresponding URID entries that resolve the reference to a resource that will be protected. A user will identify himself at home institute X. The Shibboleth components will exchange the necessary information so that the resource managers can decide based on the information in the authorization records whether the user can access the resource.

following functionality:

(1) Browsing. This is similar to clicking through a local file system where the directories are replaced by linguistically relevant groupings (sub-corpora). The approach is aimed at users familiar with or quickly able to grasp the underlying logical organization. A component allowing geographic browsing is also available.

(2) Structured search over the whole domain as well as within selected parts of it. With this type of search every metadata element can be addressed individually and the search for different elements can be combined into one query. Queries can be formulated with high precision required by research interests. Yet, the user has to know the terminology used by the metadata set in order to achieve a high recall. Furthermore, structured search is restricted to elements with closed or open vocabularies and does not cover elements with free text.

(3) Unstructured search over the whole domain or selected parts of it. Users can enter words or regular expressions into a free text field (Google-like). Any metadata element including the free text descriptions that contains matching strings will produce a hit. The recall with this method can be expected to be higher compared with structured search however, the precision will be poor.

4.2. Versioning of Resources.

The "stable identifier" issue addressed in 3.2 makes no sense if the resource itself is modified. Therefore, the original resource should never be deleted from an archive and always be accessible (although it need not be immediately). Also when we have a reference to a resource, we would like to be able to have access to older and newer versions if they exist. So when new resources are put into the archive and the depositor specifies they are to replace existing ones, the old resources are to be suitably marked and moved to the archive's "attic".

Discussions on the visibility in views on the archive of the old versions are complicated, but for the moment we have decided on allowing only access to older versions on the basis of a direct reference to it or via a reference to another version of it. This divides the "viewable" archive in two dimensions: (1) the set of all latest versions of all objects in the archive and (2) on the basis of a selected archive object we have access to its older versions.

4.3. Access Management System

Needed is also an efficient way to generate the authorization records for resources of whole corpora at once. Such a system should also allow archive management to delegate this task of setting access permissions to the depositor of the resource or somebody else responsible for the corpus.

At the MPI such a system is currently available although not yet integrated with Shibboleth and HS. This access management system is not DAM-LR prescribed and every partner archive can choose to implement its own version.

5. Conclusions

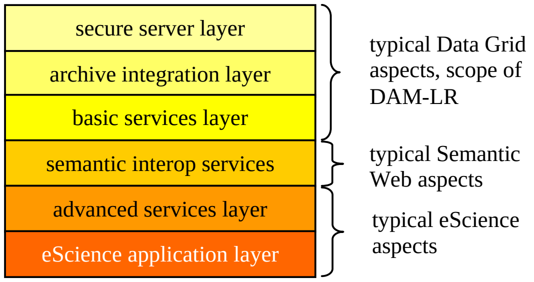
secure server layer
archive integration layer
basic services layer
semantic interop services
advanced services layer
eScience application layer
typical Data Grid aspects, scope of DAM-LR
typical Semantic Web aspects
typical eScience aspects

Figure 2 indicates the typical layer hierarchy where Grid solutions take care of typical integration aspects, Semantic Web solutions address the problems associated with interoperability in particular at the semantic level and eScience solutions provide advanced applications such as semantic weaving and web-based collaboration on top of the other layers.

The DAM-LR project is an excellent test-bed for integration and sharing technologies for the Language Resource domain and even beyond for the humanities. Also the project partners are convinced that archive federations are an essential step on the way to realize an eScience scenario for linguistics and the humanities as is indicated in figure 2. Federations will be an utterly important part of a research infrastructure that will lend services not only to linguists in the broad sense, but also to other disciplines in the humanities. They will also link up to archives that house for example ethnological, historical resources and many others. Due to the virtual integration aspect of archives it is obvious that federations will bring an added value to the researcher.

Since DAM-LR is – as far as we know – the first project in the humanities that applies Grid-type of technology on a supra-national scale, it will have a great impact on establishing stable research infrastructures in the humanities. Therefore, we feel that it is important that all DAM-LR documents be made openly available and a training program be created to actively inform other interested parties. Also DAM-LR was purposefully setup as a small project with initially a few partners, but, given the architectural simplicity of the solution found, it is our intention to scale DAM-LR up to possibly up to 20 European partners if enough interested resource archives can be found that can offer well organized documented resources.

6. References

[1] live archives, http://www.mpi.nl/dam-lr/live-archives
[2]GRID forum, http://www.gridforum.org
[3] DAM-LR project, http://www.mpi.nl/DAM-LR/
[3] GTK, http://www.globus. org/
[4] PURL, http://www.purl.org
[5[HS, http://www.handle.net
[6] http://www.mpi.nl/IMDI
[7] Wittenburg, P., Peters, W., Broeder, D. (2002). Metadata Proposals for Corpora and Lexica. In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on

Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326

[8] OAI/PMH http://www.openarchives.org/OAI/openarchivesprotocol.html

[9] CNRI. http://www.cnri.net

[10] TACAR. http://www.tacar.org/

[11] EUGRID, http://www.eugridpma.org/

[12] PKI, http://www.pki-page.org

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

[14] http://www.science.uva.nl/research/air/projects/aaa

[15] http://www.openldap.org