

A Grid of Language Resource Repositories

Daan Broeder, Remco van Veenendaal, David Nathan, Sven Strömqvist
MPI for Psycholinguistics, Institute for Dutch Lexicology, SOAS University of London, Lund University
daan.broeder@mpi.nl, veenendaal@inl.nl, djn@soas.ac.uk, sven.stromqvist@ling.lu.se

Abstract

The DAM-LR (Distributed Access Management for Language Resources) 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. It is the basis for establishing a research infrastructure for language resources which will finally enable eHumanities. Currently, the complete architecture is designed based on a few well-known components and some components have already been tested. Based on the technological insights gathered and due to discussions within the international DELAMAN (Digital Endangered Languages and Music Archives Network) network the ethical and organizational basis for such a federation is defined.

1. Introduction

There is a general trend towards the centralized storage of language resources in digital repositories, which we call here language resource archives (LRA). An emerging number of such archives can be seen operating in the areas of field and documentary linguistics and involving institutions such as MPI¹, SOAS², AILLA³, Paradisec⁴ and LACITO⁵, as well as in corpus and computational linguistics where, for example, Lund Archive, BAS⁶, INL-TST⁷ and ELDA⁸ are active. We interpret the task of LRA to include not only long-term data preservation but also, importantly,

implementation of services allowing access to and enrichment of existing content. Such services are most likely to be provided via the Internet, especially since network bandwidths are set to increase and make it possible to effectively transfer audio and video streams. Old distribution models using optical disks will only be used in certain cases, such as where large corpora are required for the training and testing of stochastic models.

The Internet also has the potential to integrate fragmented resources. There is no longer any reason for researchers to be confronted by an assortment of idiosyncratic interfaces and access management mechanisms. If co-operating archives have resources for particular languages (or any other resources that researchers might wish to aggregate), they should aim to provide users with a seamless domain for search and access. Creating such a joint domain requires integration and interoperability at a number of levels:

- a common access mechanism so that users can enter the joint domain with a single identity and a single sign on;
- a unified domain of resolving unique resource identifiers;
- a common domain of deep metadata allowing users to locate individual resources and to carry out research queries;
- services allowing users to overcome structural and format differences of resources within and across archives;
- ontology mechanisms that allow users to overcome differences in labeling systems as far as possible.

The last two levels are typically discussed under the heading of the “Semantic Web” and are the subject of many projects such as LIRICS⁹ [1] and GOLD¹⁰ [2]. In this paper, however, we focus on the first three levels, which generally fall under the heading of ‘Grid computing’. Initially, Grid computing was driven by

¹ Max-Planck-Institute for Psycholinguistics Nijmegen, NL

² School of Oriental and African Studies, University of London

³ Archive of the Indigenous Languages of Latin America, Austin

⁴ Pacific And Regional Archive for Digital Sources in Endangered Cultures, University of Sydney

⁵ Laboratoire des langues et civilisations à tradition orale, CNRS Paris

⁶ Bavarian Speech Archive, University of Munich

⁷ Centrale voor Taal- en Spraaktechnologie at the Institute for Dutch Lexicology, Leiden, NL

⁸ Evaluations and Language resources Distribution Agency, Paris

⁹ Linguistic Infrastructure for Interoperable Resources and Systems, EC-funded project

¹⁰ General Ontology for Linguistic Description, E-Meld project

high performance computing challenges in natural sciences, and focused on problems such as performing large computations using a number of high performance computers in tandem. In humanities disciplines the focus is not yet on sharing computing power, but rather on virtual integration of increasingly large data repositories. Data integration appeared within the grid community as the Data Grid track [3], and, in addition, much relevant work has been carried out within the Digital Library community. In this paper we will focus on these aspects.

In the domain of language resources, two initiatives have been taken to tackle problems arising when creating a virtual domain of language resource archives:

- the international DELAMAN initiative [4] works towards defining and creating a world-wide federation; and
- the EC-funded DAM-LR project [5], in which the four LRAs involved (MPI Nijmegen, Lund, INL Netherlands, SOAS) are in the process of establishing a federated access system. At present, they have designed a complete architecture and are currently implementing it.

In this paper we discuss plans and results to date of DAM-LR, based on a description of the underlying technologies. Finally, we note that other aspects such as frameworks of trust, ethical and legal operation also need to be addressed to create an effective federation.

2. Federation Technologies

Four technological pillars are essential to the establishment of a federation of archives:

1. an integrated metadata domain that allows users to browse and search in a federation-wide metadata catalogue and to create their own work space by selecting resources from the various archives of the federation.
2. a single resource domain where each resource is identified by a unique resource identifier. This should allow for transparent access to a resource even where multiple instances are held across different federation sites.
3. users need a single identity accepted by all federation members so that users only need to authenticate themselves once in a single session in order to access resources at all members' sites
4. an authorization system is needed that allows archive managers to give federation-wide

access to users and groups that have the appropriate rights.

These pillars are based on the existence of a domain of trusted servers and services – each component has to make sure that it is provably authenticated to be the one that it claims to be. As a means of implementing such a trusted domain, the TACAR¹¹ list [6] of mutually agreed certificates was created, based on the principles of EUGridPMA¹² [7]. In this implementation, national bodies declare that they will accept certificates from each other, with a Public Key Infrastructure [8] used to sign certificates. Every federation member has to apply to their national Certificate Authority to request the status of a Registration Authority. Once recognized as a Registration Authority, sites can request certificates that will be accepted within the EUGridPMA domain.

With respect to metadata interoperability the IMDI¹³ metadata infrastructure [9] will be supported for browsing and searching either by using stored IMDI metadata or by creating them on the fly from a local format. IMDI was chosen because it supports not only resource discovery but also resource management which is regarded to be an essential function within federations. Several portals will be made available with full functionality of metadata browsing and searching. For harvesting two methods can be applied: the OAI PMH¹⁴ protocol [10], or harvesting of native IMDI XML metadata.

The second pillar is the creation of a unified domain of unique resource identifiers (URIDs) to provide a stable method for referencing electronic resources. There are many reasons for introducing URIDs. A URID:

- is intended to persist over time
- is independent of the resource location
- is always associated with a unique resource
- can be resolved to multiple copies at different locations

They can be compared with ISBN numbers that are used to uniquely identify published books. The federation partners need a system to create, manage and resolve URIDs. They chose the Handle System [11] which is used by many well-known institutions such as the Library of Congress. To implement URIDs using the Handle System, an institution requests a centrally specified prefix that uniquely identifies its

¹¹ TERENA Academic CA Repository

¹² European Policy Management Authority for Grid Authentication

¹³ ISLE Metadata Initiative

¹⁴ Open Archives Initiative, Protocol for Metadata Harvesting

local domain. The institution is then free to specify its own postfix system. The members discussed whether we should adopt a common syntax for the postfixes. Ultimately, it was agreed that while there is no necessity for such a unification, postfix strings should not include semantically significant components.

The federation has agreed to maintain at all sites the access rights statements for a given resource as defined by its originating member. Since URIDs point to the originator's copy of a resource, it was decided that the access authorization information is associated with URIDs, i.e., it will be stored in the Handle System records. The Handle System records will be redundantly stored at multiple sites, but the originating member will have all rights on the copies, i.e., no one else will have control about modifications etc. A view of a Handle System is shown in Figure 1.

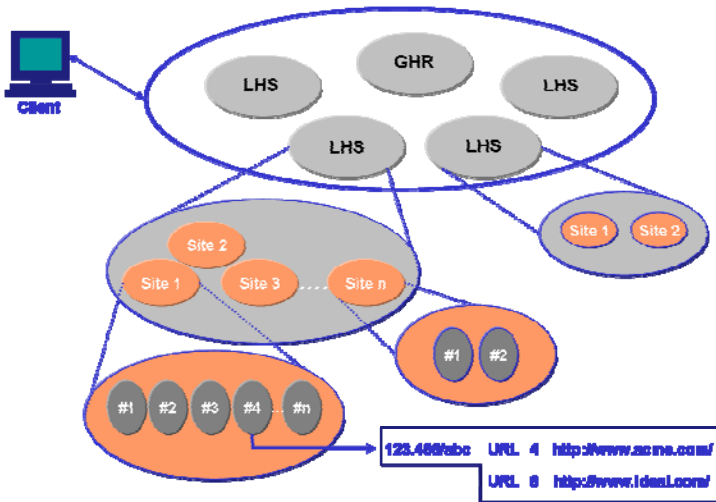


Figure 1 shows a typical Handle System scenario with a Global Handle resolver, different Local Handle Systems that can have various sites and where each site can share the job between different servers. This allows us to implement redundant services and scale up with the amount of requests to be handled.

With respect to authentication and authorization the situation is more complex. One widely used contender for implementing these, Shibboleth [12], is excellent in circumstances where users can be described by attributes such as “member of university class X” or “member of staff category Y”. The authentication of the student or staff member is left up to the home institution and the others grant access to resources based on the attributes that specify a class membership. However, for researchers operating autonomously, as will often be the case for our users, this method does not work, because authorization requires institutionally-supplied attributes that identify individuals, such as a unique ID. So, one of the attributes shared needs to identify each user uniquely.

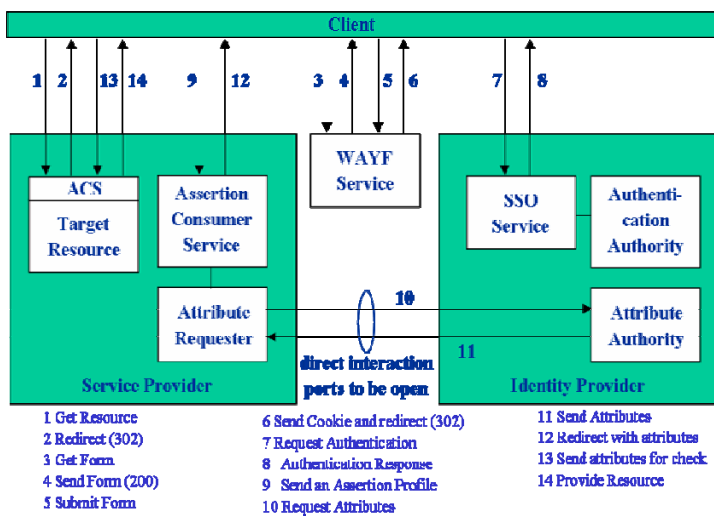


Figure 2 shows a typical scenario when using Shibboleth. All interaction is done by making use of the redirect services. The local site has to provide a suitable Access Control System and an Authentication services. Shibboleth is basically used to exchange user credentials in a safe way.

Other proposals such as using the AAA¹⁵ toolkit [13] that emerged from the Grid community [14], or using LDAP¹⁶ [15] not only for authentication but also for authorization were of interest. All these frameworks have advantages and disadvantages. Finally, the fact that Shibboleth has already received wide acceptance in the Digital Library domain influenced our choice. The interaction path for Shibboleth is shown in Figure 2.

The federation partners agreed that user management will be performed at the home site and that

¹⁵ Authentication, Authorization & Accounting

¹⁶ Lightweight Directory Access Protocol

only limited data about users will be exchanged. The prototypical system will support Open LDAP for user management since it has many useful features, it is already widely used in the academic world and it offers an interface to Shibboleth. LDAP also has the advantage of providing a simple solution to the problem of authenticates autonomous (non-institutional) users. Large institutions such as universities decide about user accounts, resources and rights at a very high level. These policies will differ from the requirements within a federation. LDAP provides a simple way to setup a local departmental LDAP service that is based on the institutional user information (filtered dynamic copy), but that also allows the department to add other users and to manage other attributes. Each federation partner is free to setup their own authentication system; however, all communicating interfaces have to be consistent across the federation.

A few components need to be added to complete the architecture. Firstly we need an Access Control System that will guard protected resources (that are in principle all accessible through ordinary HTTP requests) and force authentication and authorization via Shibboleth. The authorization records are also available as an addition to the URID record via the HS because partner archives housing a copy of the resource need to enforce the same access restrictions and so need reliable access to this information. We plan to implement the Shibboleth Identity Provider component, that will take care of authentication of users, as a TOMCAT [16] container using a JAAS¹⁷ realm. In this way it can also take care of authenticating users for related applications such as the archive's browsing and upload web-applications. The actual protecting of resources, the Shibboleth service provider, is done by the "standard" Shibboleth apache web-server module.

Secondly we need a management system that allows archive managers to efficiently manage the archive user records and set policies and permissions for their access to the resources. Perhaps a third component would be one where (new) users of the archive can request for access to archive resources and if no archive manager intervention is required (as in case of

agreeing with a code of conduct) access would be automatically granted.

Figure 3 shows a complete architecture for a typical scenario where a user wants to access a single resources using a web browser. For more complex access scenarios which involve using applications such as ANNEX [17] and LEXUS [18] to access multiple files, adaptations have to be carried out to support

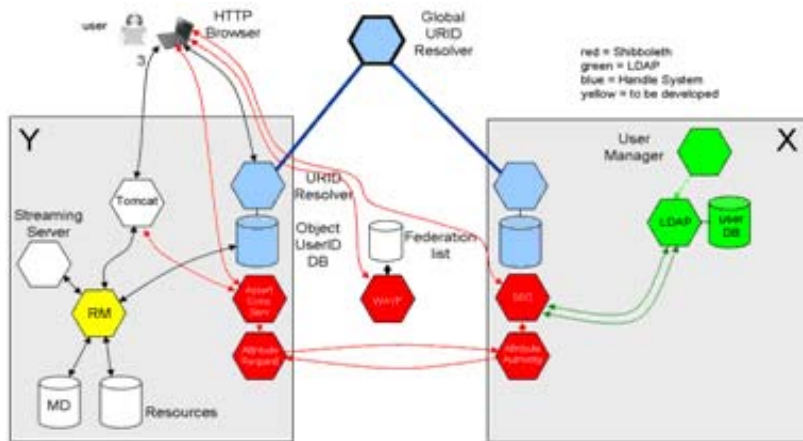


Figure 3 shows the architecture of accessing one resource from repository Y by a user homed at institution and working with a normal HTTP browser. The figure also indicates the interaction between the different components involved. It has to be noted here that a streaming server has to be integrated to support access to media streams.

working on a basket of resources which may come from different repositories.

It should be noted that the DAM-LR project does not include the exchange of resources, although this is finally intended by the partners. To achieve this more components will have to be added to ensure that distributed copies remain identical and that updates are exchanged. Different schemes were investigated. To carry out some tests in DAM-LR we will ensure that updates need to be made at the originating institution and that dynamic protocols allow new versions to be propagated to the other sites. Maintaining checksums together with the URIDs will allow us to trace efficiently whether content has been changed.

3. Federation of Archives

So far we have described a number of essential aspects in implementing a system for presenting unified access to resources across a number of LRAs. However, such a federated system has to be built on more than just technology. A federation has to be based on factors such as:

¹⁷ Java Authentication and Authorization Service

- a shared mission to provide integrated services;
- mutual trust that the participating LRAs follow agreed operating rules, such as about the management of user accounts, and respecting access conditions formulated by the originating LRA;
- ethical and legal rules in regard to exchanging and disseminating data;
- practical definitions such as the user attributes to be held and exchanged.

Access conditions are an important aspect of managing language resources. It is not feasible to formulate access conditions on a federation-wide basis because some conditions are associated with individual archive objects; others are a consequence of local archive policies. Nevertheless, the partners are formulating the range of forms that members' access conditions can take and how these are to be encoded, since some of them will need to be implemented as part of the authorization process for portal access. To the extent possible we will identify general principles governing access to archival objects that can be shared by all partners. On the other hand, *access agreement* - the process whereby users agree to abide by a condition statement and which thereby creates a contractual agreement - is an area which should reflect the partners' joint and mutual interests and responsibilities. A uniform access agreement is an important manifestation of the federal nature of the partners, and provides users with consistency at the point where legal and ethical matters apply. Therefore, the access process reflects the federal organization which provides it: in a typical usage scenario the user starts at a single, site-aggregating portal, progresses to view resources that exist across different sites, but finally makes a binding access agreement that is uniform for all resources.

4. Conclusions

The DAM-LR partners are convinced that archive federations are essential on the way for realizing an eScience scenario for linguistics. In doing so federations of language resource archives form an utterly important part of a research infrastructure that will lend services not only to linguists in the wide sense, but also to a wide number of 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 of archives it is obvious that federations will bring an added value to the researcher.

We see DAM-LR not only as a test-bed for the integration technology, but also as a way to establish a usable robust domain of services and servers that may be extended by other archives joining later. 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. Even beyond this we can say that due to our discussions on an international scale within the DELAMAN framework the experience gathered in DAM-LR will be very influential for proposals in other countries such as the US and Australia and for initiatives that cross the European borders. Already now we are in discussion with centers overseas to become LRA and to join the federation.

5. References

- [1] <http://lirics.loria.fr/>
- [2] <http://www.linguistics-ontology.org/tools.html>
- [3] www.grid.ro/workshop/documente/ppt/Fabrizio_Gagliardi_RoGrid_April_02.ppt
- [4] <http://www.delaman.org>
- [5] <http://www.mpi.nl/DAM-LR/>
- [6] <http://www.tacar.org/>
- [7] <http://www.eugridpma.org/>
- [8] <http://www.pki-page.org>
- [9] <http://www.mpi.nl/IMDI/>
- [10] <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [11] <http://www.handle.net>
- [12] <http://shibboleth.internet2.edu>
- [13] <http://www.science.uva.nl/research/air/projects/aaa>
- [14] <http://www.gridforum.org>
- [15] <http://www.openldap.org>
- [16] <http://tomcat.apache.org/>
- [17] <http://www.mpi.nl/annex>
- [18] <http://www.mpi.nl/lexus>