

Architecture for Distributed Language Resource Management and Archiving

Peter Wittenburg, Heidi Johnson*, Markus Buchhorn^, Hennie Brugman, Daan Broeder

Max-Planck-Institute for Psycholinguistics, *AILLA Houston, ^ANU Canberra
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.wittenburg@mpi.nl

Abstract

An architecture is presented that provides an integrated framework for managing, archiving and accessing language resources. This architecture was discussed in the DELAMAN network – a world-wide network of archives holding material about endangered languages. Such a framework will be built upon a metadata infrastructure, a mechanism to resolve unique resource identifiers, user and access rights management components. These components are closely related and have to be based on redundant and distributed services. For all these components existing middleware seems to be available, however, it has to be checked how they can interact with each other.

Introduction

At LREC 2000 in Athens a first workshop was organized by two of the authors dealing with open metadata for language resources as a means for easy resource discovery. At this workshop the IMDI (ISLE Metadata Initiative) concepts were presented. At an international meeting about metadata for language resources in Nijmegen in November 2000 this initiative presented its first infrastructure¹. Almost in parallel the OLAC initiative was started in the US by S. Bird and G. Simons. At a meeting at University of Pennsylvania in December 2000 they presented the basic ideas of OLAC².

At the LREC 2002 meeting first complete metadata infrastructures were presented by the two initiatives and recently both came up with new and improved versions of their metadata sets and their tools. We can say that both frameworks have reached a mature and stable state which is very important for the community. Also at LREC 2002 a working item about metadata for language resources was created within ISO TC37/SC4³ to create an ISO standard in this area. Meantime a proposal was worked out and is in the process of being discussed. It is based upon the excellent work that was done in the two initiatives and for example also the work within TEI⁴.

The approaches of the two initiatives are complementary in a certain sense. While OLAC started from the Dublin Core idea of creating a platform for semantic interoperability by defining only a few categories (in fact the 15 DC elements plus a language element indicating the language a resource is about), the IMDI initiative took another approach. Unbiased discussions were started with corpus linguists and language engineers to formulate their requirements. Also the TEI proposal for detailed language resource description was analyzed to come up with a first IMDI version. Therefore, the IMDI was richer and more

detailed, has flexibility in it allowing researchers to extend the core metadata set.

OLAC took the position of the well-known Open Archives Initiative to act as a metadata service provider, i.e. it focuses on harvesting metadata records and allowing searches across all harvested metadata records. IMDI took a different view since it did not only want to use metadata for discovery purposes, but also for managing the continuously growing language resource collections. At the MPI for example a language resource corpus covering more than 30.000 sessions has to be maintained which would not be possible without useful metadata.

Several other institutions and researcher teams started to create and integrate open metadata to improve the visibility of language resources. In this respect we can refer for example to the ECHO⁵, INTERA⁶ and DOBES⁷ projects. Despite the results of the ENABLER⁸ project showing that the majority of language resources still is not visible we can speak about an emerging critical mass of visible language resources.

Based on this progress and the growing online language resource collections in particular in the area of endangered languages DELAMAN (Digital Endangered Languages and Music Archives Network) was founded. It was meant to go beyond the visibility of language resources through interoperable and distributed metadata descriptions. The ideas are driven by the fact that the involved archives house data that is interesting for a researcher community that is distributed world-wide, that for example members of the indigenous communities are not interested in archive boundaries, but want to access “their” material and that only world-wide distribution of the material will guarantee long-term preservation in the digital era.

DELAMAN

The DELAMAN network currently comprises the following archives:

¹ All IMDI related work and also the mentioned events are documented at two web-sites: <http://www.mpi.nl/ISLE> and <http://www.mpi.nl/IMDI>.

² Open Language Archives Community: <http://www.language-archives.org>

³ ISO TC37/SC4: <http://www.tc37sc4.org>

⁴ Text Encoding Initiative: <http://www.tei-c.org>

⁵ European Cultural Heritage Online: <http://www.mpi.nl/echo>

⁶ Integrated European Language Resource Domain:

⁷ Documentation of Endangered Languages:

<http://www.mpi.nl/DOBES>

⁸ <http://www.enabler-network.org>

- AILLA (Archive of the Indigenous Languages of Latin America) located in Austin, Texas
- Alaska Native Language Center Archives
- DOBES Archive (Documentation of Endangered Languages Programm) located in Nijmegen, Netherlands
- E-MELD Project located in Ypsilanti, Michigan
- ELAR (Endangered Language Archive of the Hans Rausing Endangered Languages Project) located in London
- LACITO (LANGUES *et* CIVILISATIONS à TRADITION ORALE) located in Paris
- MPI Archive located in Nijmegen, Netherlands
- PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) located in Sydney

The DELAMAN initiators agreed that a number of issues have to be tackled at a world-wide level where the individual archives have to collaborate. In this paper we want to focus on those topics that are addressing the need to come to an infrastructure for a distributed language resource management and archiving.

Three goals were seen as being central: (1) long-term preservation strategies for the stored human heritage material, (2) an efficient rights management system integrating the archives to one virtual one and (3) the formation of a world-wide user group where users have just one identity to access the resources.

Long-term Preservation

Current storage media is vulnerable and has a very restricted life-time of a few years. This may indicate that our current magnetic and optic storage technologies are not suitable for long-term preservation purposes. However, there is no other technology that allows us to store several terabytes of data that also is changing frequently. The short life-time can be overcome by regular migration to new storage media and this is what computer centers are used to optimize for years now. The vulnerability has several causes: (1) The technology itself is not safe compared for example to paper or even clay tablets as they were used by the Sumerians for example. (2) There are several possible emergency scenarios ranging from environmental to political instabilities. It is well-known that only wide-scale distribution can overcome this problem in the assumption that one or a few of the copies will survive and migrate between storage systems and technology.

Therefore, the DELAMAN partners have the goal to lay the ground for sharing the data at a world-wide level. It is utterly valuable data about human heritage what the archives are storing so it is worth using modern networks with their growing bandwidth to

exchange the data dynamically. It is expected that within the coming 5 years all problems ranging from technological to ethical ones can be solved.

User and Access Management

The DELAMAN partners are aware of the fact that the user community interested in the data they are archiving is distributed world-wide. These are researchers, journalists, school and university classes and even members of the indigenous communities who are interested in comparing different language types at different linguistic levels, who want to carry out deep linguistic studies, who want to listen to the voices of their ancestors and many other purposes.

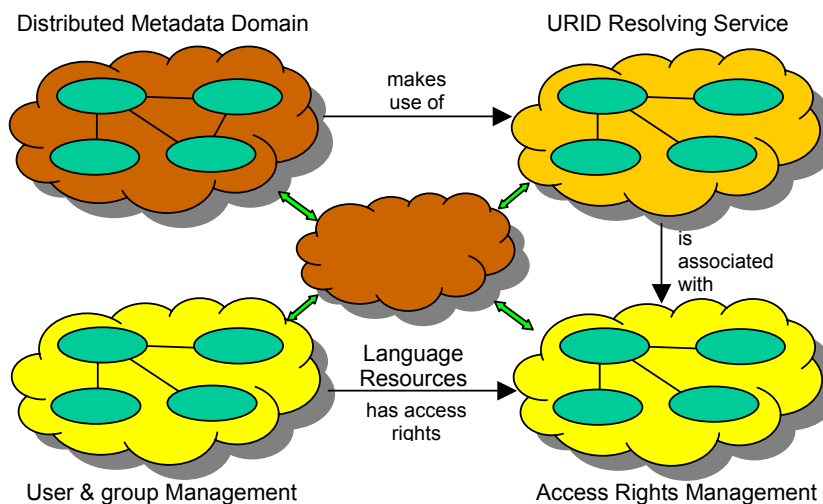
Currently, the situation is so that archives A and B both may have data about a language X leading to distinct sub-corpora XA and XB. The users can view only one of the sub-corpora per time which is caused by a number of reasons: (1) Each user currently has many identities – in general one per archive. (2) Each archive has different access management policies and types of interfaces to the data. (3) The interoperability between resources is limited both at syntactic and semantic level.

DELAMAN wants to address the first two points knowing that a lot is currently being done to increase the interoperability at the syntactic and encoding level (XML, UNICODE) and at the semantic level (ISO Data Category Registry, RDF(S), OWL). It is intended to create one integrated user domain for all participating archives such that a user only has one identity.

Further, it is the intention to have one domain where archivists and data managers can define groups of users and associate access rights with them. Together with a distributed and integrated metadata domain of language resource descriptions this would create the intended integrated access domain.

4 Pillars

The implementation of the ideas presented by Wittenburg (2003) and then discussed within DELAMAN is based on four essential pillars as is indicated in Figure 1. The metadata descriptions describe the characteristics of the



language resource objects for discovery and management purposes. This means that they can act as the glue to bundle relationships in a natural way. The IMDI framework for example allows us to do so. The metadata descriptions have to point to the real resources.

However, the metadata descriptions should not be used to resolve unique identifiers to physical paths and they should not be misused to store all physical paths. The metadata descriptions have to point to the abstract objects which are represented by unique resource identifiers. It is the URID resolving mechanism that translates from the URID to the different physical instances. In doing so we create just one place where changes have to be maintained that have to do with copying and migrating of data.

Access rights are typically not associated with the instances but with the object. Independent of where the instances are stored the same group of users will have the rights to access them. There may be differences for internal services, but they are out of the scope of DELAMAN.

Finally, there is one distributed user and group domain such that all managers that have the rights to define access rights can see the same set of users. On the other hand the users have one identity when navigating through the integrated domain of language resources and accessing them.

The four briefly introduced services are the pillars of what is called a virtually integrated domain of language resources without eliminating the responsibility of every archive to define its policies. It is especially useful to implement a unified access management in a domain of distributed resources.

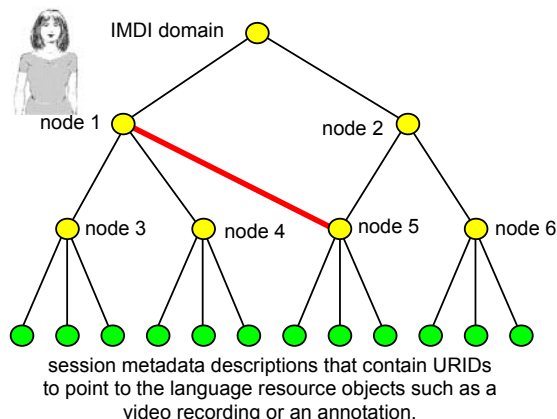
Since the users on the one hand will become very dependent on the functioning of these pillars, each of them has to be realized with high availability in mind. If the URID resolving mechanism would rely on one single server it can be easily understood which kind of disaster would be created if this server would be down for some time period. On the other hand the implementation with help of a set of synchronized services would give the user a much higher availability than it can be offered now by one institution.

Metadata Descriptions

As indicated the domain of linked metadata descriptions is the one in which users navigate when they want to discover useful resources and in which resource managers will operate to efficiently carry out typical management tasks such as changing access rights or copying a sub-corpus.

So metadata descriptions have to serve more functions in this case than being harvested to facilitate searching. The canonical trees in the IMDI domain for example are used now to specify access rights in a very efficient way, while maintaining the different project responsibilities. This is indicated in Figure 2.

The structure shows a distributed domain of metadata descriptions where the sub-tree “node 1” can reside on a different server as the sub-tree “node 2”. The tree could exist of several layers of abstraction representing institutions, languages, researchers/projects, etc. Here we can only show two layers and let us assume that the “node 1” and “node 2” sub-trees belong to different projects, i.e. different authorities that can lend access.



We can now imagine that the data manager of the “node 1” sub-tree can select a certain node in the tree and define with one statement, that all resources of a certain type such as all videos under this node should be accessible by a certain group of users. This is very efficient, especially if the manager can delegate responsibilities to other persons. Since users are allowed to create their own virtual IMDI metadata domains by linking nodes in their own way and by creating additional nodes, one has to take care that these are not being used when inheriting access rights. The red line indicates such a new link created by some user to link two nodes from different responsibility domains.

At the MPI such a scheme has been realized now making use of the IMDI metadata infrastructure. So the concept of abstract nodes in a corpus hierarchy can be effectively used for this type of management activities.

Such a distributed metadata domain looks very similar to a distributed file system such as the Andrew File System. The big difference is given to the users. The metadata files contain linguistic type of descriptions, i.e. meaningful and interpretable data.

Unique Resource Identifiers

A consequence of the migration and copying efforts is that there will be several instances of the language resources. Therefore, it seems to be necessary to differentiate between an abstract object and its instances. This also means that a mechanism has to be introduced that identifies the objects. We would like to introduce, therefore, the concept of unique resource identifiers (URIDs) that is well-known in the archive and library world. URIDs represent the abstract object and all characterizations such as metadata and attributes such as access rights have to be linked to it. It is the URID resolving mechanism that translates to the physical paths of the resources. So metadata descriptions and URIDs are

complementary: while URIDs are arbitrary codes that represent the knowledge about the locations of the instances, the metadata descriptions describe the objects.

Is there middleware that can be used? There are some disciplines where the need to introduce URIDs became apparent already at an earlier stage such as the DOI initiative⁹. For us there are two interesting solutions. The PURL (Persistent URL)¹⁰ solution offers an interesting option, but its functioning is dependent on a central service which is not acceptable. The Handle System¹¹ solves this disadvantage of PURLs in so far that it allows to have a multitude of services and redundancy. Further, it allows authorities to define their own encoding schemes for URIDs and to associate information such as physical paths with each entry. A URID has the basic distinction between an authority indication and a unique code created by that authority.

By allowing redundant services the Handle System provides high availability and improves the resolving performance. So, first tests with an installation seem to indicate that there is already very useful middleware for URID resolving which is also optimized with respect to the speed of operation. Compliance to an emerging OpenURL standard is guaranteed since the creators are active members of the corresponding discussions.

User/Access Management

As indicated, the goal is to come to an integrated and unified user and access management system for the interested institutions within DELAMAN and beyond. Also this type of service requires high availability to be accepted by the users. For the managers it must offer simplifications, since it can be expected that many users will request access to the resources. It was already said that the access rights have to be associated with the objects represented by the URIDs and not the individual instances. So whatever system is chosen it has to be integrated with the described linked metadata domain and with the URID resolving mechanism.

The solution must fulfill a number of requirements such that managers can delegate “manager” rights to other persons for sub-corpora, that users can be clustered into groups to which access rights are given, that users of course can be members of several groups, that access rights can be given for special purposes and limited periods, that certain declarations with legal and/or ethical content have to be signed and others more.

There are two candidates for useful middleware that comes close to what is needed. The LDAP system¹² allows to setup a hierarchical nevertheless distributed system to primarily administer users. Its middleware built on top of databases is optimized to support fast authentication and to add other type of information. So it seems to be a perfect solution for managing users,

however, it is not a system that is ready to do access rights management.

The Shibboleth project¹³ has a complementary but partly overlapping focus. It is developing an open, standards-based solution to the needs for organizations to exchange information about their users in a secure, and privacy-preserving manner. The purpose of such an exchange is to determine if a person has the permissions to access a given resource based on membership of institutions, groups etc. In this scenario users are accepted based on attribute assertions provided by his origin campus or home site, who are responsible for authentication. Access control decisions by the target site are based on the interaction about these attribute assertions. Therefore, Shibboleth seems to be an excellent candidate implementing the access management pillar.

Summary

We have described the goal of DELAMAN to come to an integrated resources management domain that covers the 4 essential pillars (1) metadata descriptions, (2) URID resolving mechanisms, (3) user management and (4) access rights management. The solutions chosen for these 4 pillars have to work together to achieve the level of integration that is required by users and managers and all services have to offer a high availability and fast performance to be accepted.

For each of these pillars we can identify solutions that either serve the needs or that come close to what is intended. For the metadata solution the IMDI infrastructure is an acceptable framework, for resolving URIDs to physical paths the Handle System seems to be an excellent candidate and for the user and access management LDAP and Shibboleth middleware are good candidates.

It will be investigated in the coming months what Shibboleth offers in detail and how it can be interfaced with the other services. If our expectations are met we hope that we will be able to create a first version of an integrated solution in the next year.

References

Wittenburg, P. (2003). Archiving Strategies for Multimedia Language Documentation. Paradisec Workshop about Speech Archiving. Sydney, October 2003

⁹ Document Object Identifier: <http://www.doi.org>

¹⁰ Persistent URL: <http://purl.oclc.org>

¹¹ The Handle System: <http://www.handle.net/>

¹² LDAP: <http://www.openldap.org/>

¹³ Shibboleth: <http://shibboleth.internet2.edu/>