

Deliverable 2.2

Local Prototype Final Report

DAM-LR

011841

Distributed Access Management
for
Language Resources

implemented as
Specific Support Action

Contract Number: *011841*

Project Coordinator: Peter Wittenburg

Project Web-Site: www.mpi.nl/dam-lr/

Deliverable: D2.2

Date: 20.2.2006

Content

PREFACE	3
1. GENERAL	3
2. LOCAL PROTOTYPE	4
2.1 METADATA LAYER.....	4
2.2 PHYSICAL LAYER	4
2.3 ACCESS MANAGEMENT	4
2.4 INGESTING RESOURCES	5
2.5 ARCHIVE MANAGEMENT	5
APPENDIX A: DAM-LR LREC WORKSHOP CONTRIBUTION.....	6
APPENDIX B: DAM-LR LREC CONTRIBUTION.....	11
APPENDIX C: MPI LREC CONTRIBUTION	16
APPENDIX D: MPI LREC CONTRIBUTION	21
APPENDIX E: MPI LREC CONTRIBUTION	27

Preface

The prototype was developed in full accordance with the specifications, therefore we have chosen to re-use the chapter organization from deliverable 2.1 where possible. A number of chapters are of general nature, therefore we will omit them. In particular the old chapter 4 now becomes chapter 2. For every requirement on the original specification list we will report about the state.

1. General

In our specification report we sketched the framework within which DAM-LR will run. It is not the task of this report to repeat all statements. Instead we can refer to a number of activities that confirm our statements.

The Language Resource Archives are growing in all major centers and new archives have been established in several countries. At the MPI, to take just one example, the size of the archive is approaching the 20 Terabyte limit which means an increase of more than 50 % per year. This increase can also be described in terms of archived objects which is close to crossing the boundaries of 200.000. This trend can be seen in various centers that house language resources.

The principles that were described in the specification report are confirmed by the OAIS model and by the Live Archives flyer which was created by the DAM-LR group and which received a lot of support from various resource centers in Europe.

The first archive contributions were deposited at the MPI which already provided archiving services for external researchers and projects. The large Dutch Bilingualism Database is currently being ingested and two smaller contributions were submitted by individuals (documentation material about Narrangansett and Guenche languages). At other DAM-LR partner institutions similar activities take place.

The DAM-LR goals are implicitly very much supported by the ESFRI Research Infrastructure roadmap initiative. The DAM-LR partners are part of the CLARIN research initiative and it is obvious that

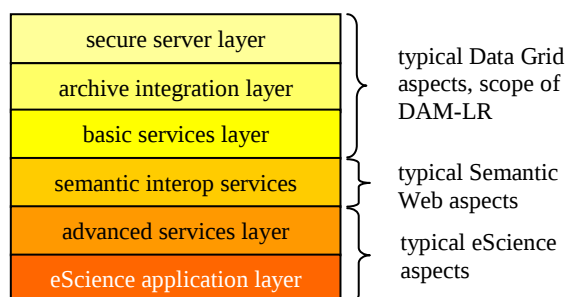


Figure 1 indicates the typical layer hierarchy where Grid solutions take care of typical integration aspects, Semantic Web solutions address the problems associated with interoperability in particular at the semantic level and eScience solutions provide advanced applications such as semantic weaving and web-based collaboration on top of the other layers.

- research infrastructures will enable eHumanities
- language resource archives are important components in research infrastructures
- DAM-LR covers the Grid layer of research infrastructures

These relations are sketched in figure 1 which is taken from a workshop paper to be given at LREC 2006 (see appendix A). DAM-LR is located at the level of organizational interoperability, i.e. it creates

- one integrated metadata domain based on the IMDI standard where the users can browse and search to discover useful resources
- one domain of unique identifiers for all resources covered
- one domain of user identity so that they can operate in the archives with a single sign-on
- one point of authorization-granting so that for these user identities access authorization to the language resources is defined, independent of where the resources are offered.
- all is based on a system of trusted servers and services.

All issues have been addressed by DAM-LR and their relevance for the local prototype, a reference installation at the MPI for Psycholinguistics, were checked in detail.

2. Local Prototype

The local prototype largely developed at the MPI for Psycholinguistics already fulfills most of the challenging criteria. For more details we refer to the papers that we will give at LREC which are included as appendices.

2.1 Metadata Layer

Requirement	Status
All archive resources are described by IMDI metadata,	done – although the degree of usage of the IMDI elements varies (appendix A).
All IMDI descriptions must be open and accessible as XML files.	done
Structured and unstructured metadata searches within the IMDI domain have to be possible.	done
Searches with the help of search engines such as Google have to be supported.	done
The IMDI resources have to be offered as OLAC/DC records according to the OAI PMH protocol to make them searchable via OLAC service providers.	done – although for OLAC special MD descriptions at “corpus” level had to be created
The IMDI files must be linkable into a unified domain that organizes resources into logical bundles that supports browsing and enables resource management.	Done
It must be easy to register and integrate new IMDI-based repositories into the IMDI domain.	Done
It must be easy to setup an IMDI portal.	done – latest example Lund University
It must be easy for users to access resources via the metadata descriptions when they are authorized.	Done

2.2 Physical Layer

Requirement	Status
The physical storage must be transparent to the user, i.e. the user should not have to deal with servers, disks etc.	Done
The physical location of the resources should be easily modifiable without causing problems for the users.	Done
The organization of the archive should allow for the copying of whole and sub-parts of the archive to new archives to support long-term preservation and redundant access paths.	Done
Each archived resource has to be identified by a unique resource identifier (URID). The metadata descriptions have to contain URIDs to refer to the resources.	In progress – the Handle System is installed, an architecture was designed and components are in a test phase
The storage concept must be such that several copies of all resources can be generated automatically and location resolving can be carried out.	several copies are generated – this component also has to be linked up with the URID database
The stored resources have to be in archivable formats and directly accessible for authorized people.	by far the most of them are – but there are always resources in some formats that have to be integrated for archival reasons

2.3 Access Management

Requirement	Status
The access management system must support the definition of policies (declaration of code of conducts, usage, processes etc) and rights.	done

The access management system must support the specification of usages and temporary tickets associated with these usages.	in progress – a whole request handling system is currently being implemented on top of the access management system
The access management system must support efficient electronic operations via web interfaces and a delegation mechanism to allow resource owners to define access policies from remote sites.	done
Access policy specifications must be based on the metadata layer, i.e. the physical layer is transparent to the definer and the specifications are independent of the physical location of the resources.	done
It must be possible to specify domains of authority in the metadata layer.	done
The delegation of rights must be possible.	done

2.4 Ingesting Resources

Requirement	Status
The possibility to integrate new resources or update existing resources into the language resource archive has to be controlled by an upload system that ensures that its coherence and consistency is guaranteed.	with LAMUS such a system is available (appendices)
The user has to be provided with a workspace mechanism which allows him/her to arrange the data and test its compliance until it is ready for integration.	done
The upload system has to be equipped with a configurable list of permitted file types and where possible with format checkers. In particular for complex resource types, dependent types must be indicated, some being required such as a schema.	done – although for complex types an extension has to be developed
The upload must support the definition and integration of an upload node in the existing archive, archive structure, metadata descriptions and resources. It must support the proper linking of these elements.	done
Versioning must be done in the case of integrating new versions.	in progress - a smart versioning system is currently being developed

2.5 Archive Management

Several checks are already integrated into LAMUS, while others were developed as stand-alone tools. However, they are designed so that they can be integrated into LAMUS step-wise.

Requirement	Status
copying and moving data while retaining the correctness of the archive's organizational links	Done
checking the consistency of all links in the archive and modifying them where necessary	Done
checking the format and technical encoding correctness of all resources where possible	done – where parsers are available
automatically generating additional resource types for presentation purposes such as MP3, MPEG4, etc	Done
creating different types of statistics	done
the possibility of removing sub-parts of the archive which is the most dangerous operation and which therefore has to be guided	possible – but only for a senior archive manager

Appendix A: DAM-LR LREC Workshop contribution

Integrated Services for the Language Resource Domain

Daan Broeder, Peter Wittenburg, Alex Klassmann, Freddy 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 from 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

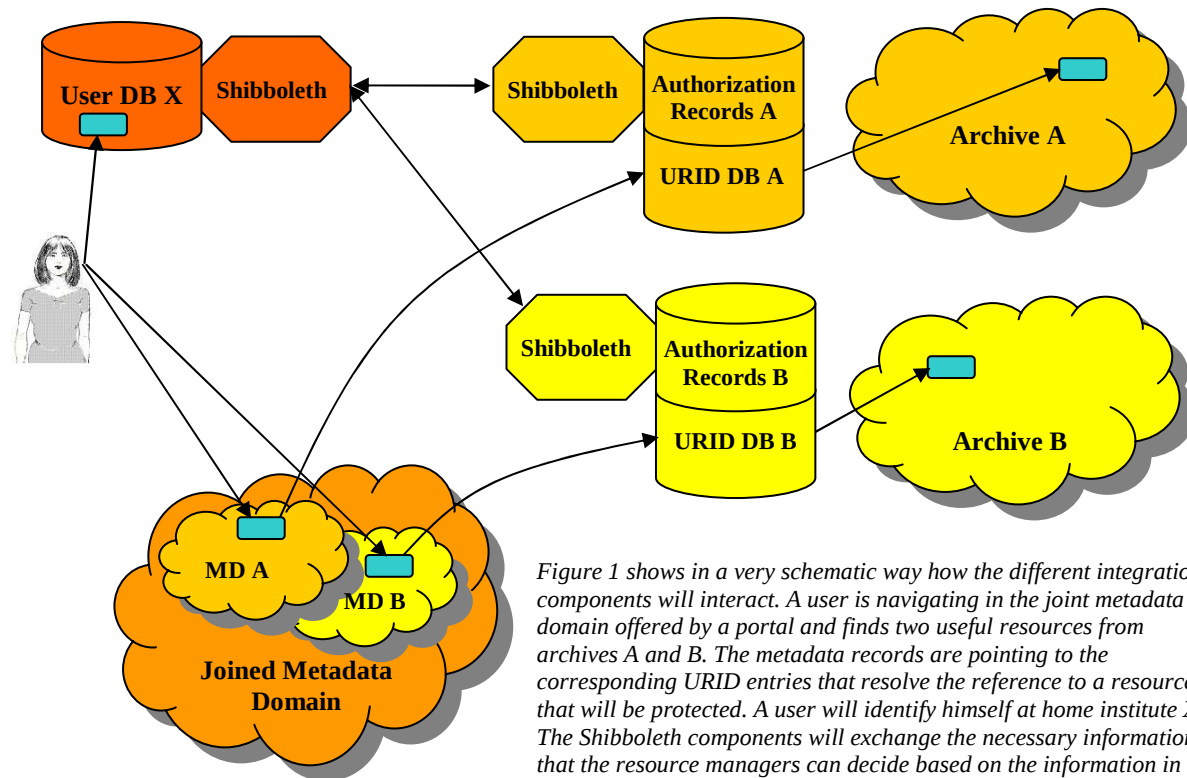


Figure 1 shows in a very schematic way how the different integration components will interact. A user is navigating in the joint metadata domain offered by a portal and finds two useful resources from archives A and B. The metadata records are pointing to the corresponding URID entries that resolve the 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.

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 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 proscribed and every partner archive can choose to implement its own version.

5. Conclusions

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.

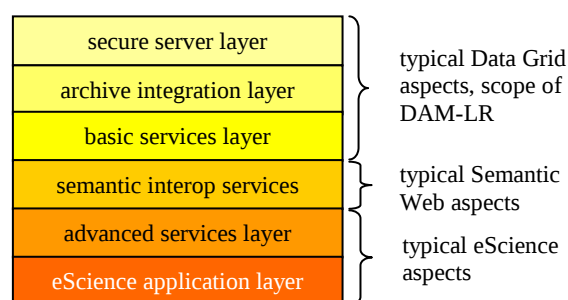


Figure 2 indicates the typical layer hierarchy where Grid solutions take care of typical integration aspects, Semantic Web solutions address the problems associated with interoperability in particular at the semantic level and eScience solutions provide advanced applications such as semantic weaving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

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>

Appendix B: DAM-LR LREC contribution

Technologies for a Federation of Language Resource Archives

**Daan Broeder, Freddy Offenga, Peter Wittenburg, Peter van der Kamp, David Nathan,
Sven Strömqvist**

MPI for Psycholinguistics, Institute for Dutch Lexicology, SOAS University of London, Lund University
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
{daan.broeder,freddy.offenga,peter.wittenburg}@mpi.nl, kamp@inl.nl, djn@soas.ac.uk,
sven.stromqvist@ling.lu.se

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. 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 are already tested. Based on the technological insights gathered and due to discussions within the international DELAMAN 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 MPI, SOAS, AILLA, Paradisec and LACITO, as well as in corpus and computational linguistics where, for example, Lund Archive, BAS, INL-TST and ELDA 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 LIRICS [1] and GOLD [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 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] (Digital Endangered Languages and Music Archives Network) works towards defining and creating a world-wide federation; and
- the EC-funded DAM-LR project [5] (Distributed Access Management for Language Resources), 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 a user only needs 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 TACAR list [6] of mutually agreed certificates was created, based on the principles of EUGridPMA [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.

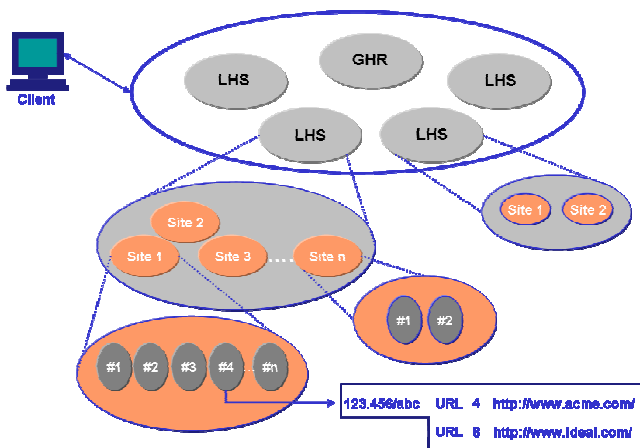


Figure 1 shows a typical Handle System scenario with a Global Handle resolver, different Local Handle Systems that can have various sites and where each site can share the job between different servers. This allows us to implement redundant services and scale up with the amount of requests to be handled.

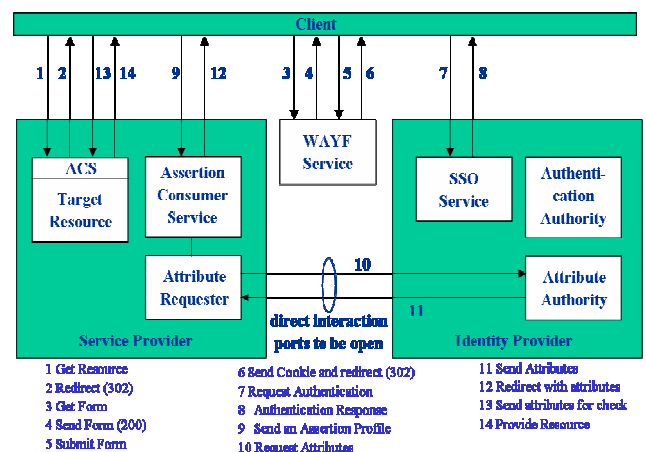


Figure 2 shows a typical scenario when using Shibboleth. All interaction is done by making use of the redirect services. The local site has to provide a suitable Access Control System and an Authentication services. Shibboleth is basically used to exchange user credentials in a safe way.

With respect to metadata interoperability the IMDI 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 PMH 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 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.

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. Other proposals such as using the AAA toolkit [13] that emerged from the Grid community [14], or using LDAP [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 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 authorizing 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 have to be added to make the architecture complete. Firstly, an Access Control System will be developed. This system will recognize requests to protected resources, redirect these requests to the Shibboleth component, and then compare the user credentials returned by Shibboleth with the appropriate records from the URID record in order to finally decide about access permission. To implement such a component we can re-use application servers such as TOMCAT [16] including components from existing software libraries. Another major component that has to be added is a management system that allows the archive manager to efficiently add new users, manipulate access policies and permissions etc. This component is specific to the DAM-LR requirements so we will develop our own, with the federation partners taking care of archive-local specific architectures.

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

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:

- 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.

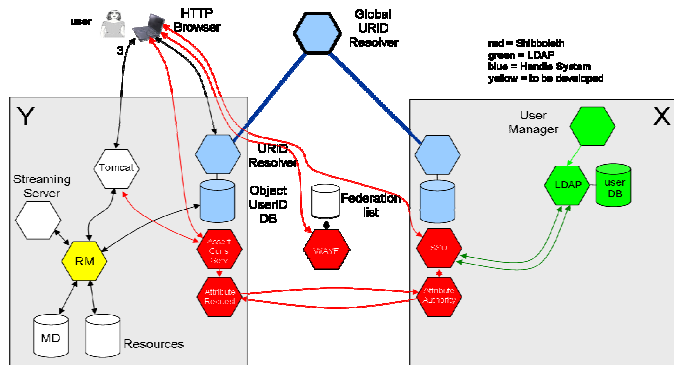


Figure 3 shows the architecture of accessing one resource from repository Y by a user homed at institution X and working with a normal HTTP browser. The figure also indicates the interaction between the different components involved. It has to be noted here that a streaming server has to be integrated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

Currently, DAM-LR includes only four LRAs. However, when DAM-LR becomes operational, we assume others may wish to join the federation; not only some other LRA institutions, but also large institutions such as universities that have many users but no shared language resources. We may need to be ready to admit some such institutions into the federation if they adhere to the same set of rules.

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 a 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>

Appendix C: MPI LREC contribution

Foundations of Modern Language Resource Archives

Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson, Laurent Romary

Max-Planck-Institute for Psycholinguistics

Wundtlaan 1, 6525 XD Nijmegen

{peter.wittenburg, daan.broeder, wolfgang.klein, stephen.levinson}@mpi.nl, laurent.romary@loria.fr

Abstract

A number of serious reasons will convince an increasing amount of researchers to store their relevant material in centers which we will call “language resource archives”. They combine the duty of taking care of long-term preservation as well as the task to give access to their material to different user groups. Access here is meant in the sense that an active interaction with the data will be made possible to support the integration of new data, new versions or commentaries of all sort. Modern Language Resource Archives will have to adhere to a number of basic principles to fulfill all requirements and they will have to be involved in federations to create joint language resource domains making it even more simple for the researchers to access the data. This paper makes an attempt to formulate the essential pillars language resource archives have to adhere to.

1. Introduction

The introduction of digital technology has fundamentally changed the ways in which we produce, store and use language resources. Traditionally, the focus was on creating publications as the result of the individual researcher's work and distributing them to share knowledge. These publications were enriched by examples the individual researcher found in his/her recordings and notes, the raw material not in general being accessible to the field. Now we have the situation that it is much easier to create digital audio or video recordings and to make them available at all steps of the scientific analysis process. In addition, large amounts of primary texts are available to the researchers via the Web and by harvesting newspaper and journal texts and digitized books. Indeed, the web is a gigantic source for language-oriented researchers and it will include an increasing amount of multimedia resources due to the preferences of the young generations. However, the web is focused on mainstream languages and language usages, i.e. it lacks most of the existing 6500 languages, many of which are highly endangered. It also lacks specific recordings where multimodal utterances are generated under controlled circumstances etc. Therefore, the creation of additional resources will remain crucial in the language research and documentation process.

This revolutionary change in data storage and retrieval possibilities has basically taken place in less than two decades and it is leading to a huge amount of primary data. We also see that the percentage of resources that are annotated, and thus which can be evaluated in terms of their scientific relevance, becomes increasingly smaller. A typical problem that traditional archives have re-occurs: repositories of digital data will contain an increasing amount of material that has not been enriched by 'value added' information in any substantial detail. The unprecedented growth of computer power and storage capacity creates the illusion for all participants in the data collection and analysis process that it is possible to manage an unlimited number of resources without additional efforts. In this paper we argue that this assumption will actually lead to the loss or inaccessibility of much of the data in a very short time.

A few examples may illustrate the utterly problematic situation. An investigation carried out by D. Schüller [1] in the aegis of an UNESCO project revealed that about 80% of our ethnologically motivated recordings about cultures and languages are endangered due to lack of care for the primary records by individuals or projects. We know that huge amounts of linguistically useful data is stored on private PCs encapsulated in some database with a high chance that this data will be lost when the PCs or the software will be retired or updated. Most of the web-sites that are used for research purposes are fragile, i.e. they will not be maintained for a longer time because funding for the project stopped or people with essential knowledge moved on. Also the lack of resource descriptions is an issue of sufficient specificity and in reusable formats is an issue. At the MPI for Psycholinguistics we had an increase of 4 TB of digital data within one year amounting to in total 15 TB. More than half of the recordings are not described by metadata, that is, there is no record of even which language is being spoken, let alone under what conditions it was recorded.

2. Language Resource Archives

Therefore, there has been an international trend to setup “centers” that are meant primarily to store all the data, which have a scientific or societal value even if they are no more than snapshots of the web documenting current language usage, or which have to be maintained simply for reference purposes. We will call these centers that store language resources, have expertise about their content and that give access services “language resource archives”. One of their main objectives is to take care of long-term preservation of the data which makes them true archives in the traditional sense. However, as the Technical Board of IASA [2] stated correctly, it is not the task of digital archives anymore to store physical objects such as tapes and CDROMs. Since digital representations can be copied without losing information and since copying can be comparatively inexpensive the situation has changed completely: it is the content and not its specific physical existence that has to be preserved. This is in particular true where we are not forced to apply lossy compression techniques and when we take care that the digital representations are complete copies.

Due to a fundamental physical law which says that we will adversely affect objects whenever we touch them, traditional archives have to impose a very restrictive access policy. In the digital domain we argue that accessing the content does not change it, which is correct if we strictly follow the stand-off annotation rules [3] and/or apply a suitable versioning system. So digital language resource archives are expected to give easy access to the material they store. This aspect is still in serious debate, but discussions within big national libraries such as the Royal Dutch Library [4] show that even such big institutions are busy adapting their business models towards more interactive access scenarios. Of course, access here is not meant in the more traditional way that institutions such as ELDA [5] provide them at this moment. They support a web catalogue and the user can ask for the distribution of the selected resource. Again, by ‘access’ we do not have in mind that resource providers offer a very restricted web-based interface with the help of which one can carry out restricted queries and access singular items.

- Based on this we can summarize what can be seen as major tasks of modern language resource archives (LRA):
- LRA have to take care of long-term preservation of the hosted data and of the stability of references.
- LRA have to offer services that allow flexible access to the data according to the needs of the potential users, and permit uploading new versions and flexibly extending them.
- LRA have to offer possibilities to enrich the data, i.e. to add new resources, commentaries and relations or update existing ones. This, may of course, not influence the archived content.
- LRA have to take care that ethical and legal constraints as well as intellectual property rights aspects are taken seriously.

LRA are service centers that address the needs of the different user groups. In the first instance, the needs of researchers have to be satisfied. In some cases also the access by native speaker communities is of high relevance. But also there are students, teachers, journalists and other groups that can be mentioned as potential user groups. Satisfying all the needs would require a whole spectrum of services that a single LRA cannot meet. Therefore, an LRA has to offer appropriate open interfaces for other service providers. The services of an LRA are amongst two extremes: very shallow, in the sense that they e.g. expose the metadata or content to simple search engines or, more deeply, rich data that offers interfaces for programmers.

3. Principles of Language Resource Archives

Based on what we have described so far, we can describe a number of principles that have to be met by modern digital Language Resource Archives. These principles, have as a corollary, that they imply requirements for technologies to be applied to them.

1. LRA have to implement a strategy for long-term preservation that includes a migration plan to new technologies and a distribution plan to create copies of the data at different locations following different protocols. This requires a kind of low-level federation, since you can only exchange sensitive data with trusted servers and organizations. This federation implies both agreements on the technology level (exchange protocols etc) as agreements in the ethical and judicial domain.

2. LRA have to adopt as much as possible widely used and open standards for all data including the metadata and relations between the resources. A conversion will be necessary towards these standards which includes structure descriptions for textual data that are compliant with generic schemas. Finally, the degree of coherence and compliance to such schemas will influence the costs of migration towards new representation formats that will emerge.

3. LRA have to differentiate between physical storage structure, which is characterized by servers, disks etc., and the linguistic archive organization, which is characterized by resource metadata and linguistically meaningful categorizations. While the first is defined by system managers and influenced by technological considerations and therefore changing frequently, the latter is determined by scientific considerations and comparatively stable. Archive management, resource discovery and usage should make use of the linguistic organization.

4. LRA have to agree on mechanisms that are able to resolve Unique Resource Identifiers (URIDs) to physical paths. Only the use of URIDs will allow us to maintain stable references and to make a distinction between an archival object and its many instances (copies) that can exist at other archives. Stable references to digital resources will become increasingly important since publications will increasingly often refer to them and are indispensable now when we want to create an interlinked domain of language resources.

5. LRA have to devise a strategy to allow selected users to upload new resources to an archive or to update existing resources without destroying the existing ones. This will require a web-based upload and management system offering work spaces and a smart versioning mechanism. It is one of the basic principles of archiving that archived data may not be touched. In the digital era this could be disastrous, since there may be references to old resources and these have to be resolved to the original objects even when new versions are available.

6. LRA must offer a powerful access management system that allows us to define access policies and offers delegation mechanisms. This is important to give depositors full control of granting access to “their” data. Relevant material will only be deposited, if the archivist declares to respect the rights of the creators and guarantees that they know that they always can access their material.

7. LRA must offer different layers of access to the data dependent on the expected user groups. This is a very problematic point since we often cannot anticipate what kind of user interface special user groups such as for example members of language communities expect. The access techniques range from geographical browsing, metadata browsing and searching to more advanced methods to access complex linguistic types such as annotated media files and multimedia lexica. Most important for an archive is to offer neutral access mechanisms that allow the user to access the individual resource without any embedding if this is required. For language technology users and to allow setting up local data centers it is often required to also offer the download of a complete sub-archive including all bundling and metadata information.

8. LRA will have to offer ontology support in the future to compensate for the linguistic encoding differences. LRA house contributions from various individuals and projects all using different terminologies to describe linguistic phenomena. Users will want to carry out for example searches across different corpora which will only work when there is smart ontology support.

9. In the future LRA will also have to offer services that allow selected users to add comments to fragments and to mark relations between them. These enrichments are part of the archive, i.e., they have to be stored in open formats including the bundling information as well. However, the original resource may not be affected.

Most – if not all - of the current repositories housing language resource data do not operate according to these principles yet. However, the pressure to do so will increase. LRA, if they are to survive in a competitive domain, will have to operate at a cost-effective level and nevertheless offer smart and stable services to the different user groups. LRA can be part of different scenarios to guarantee persistence: they can offer all services themselves, i.e., take care of redundant storage and appropriate migration strategies and access services on the one extreme end or use computer centers of libraries to take care of long-term storage and limit their own activities to providing access services. Yet we cannot rely on the services of traditional libraries and archives since they lack the knowledge about the content and have no experience with modern access scenarios as described in this paper.

4. Archive Federations

LRA will have to become members of archive federations, i.e., communities of trust and virtual integration. The term “federation” covers technological and in particular organizational and juridical aspects. In the domain of language resources we can see two related initiatives to create a federation of archives. DELAMAN [6] is an international network of archives housing endangered language and music material. Two of the major goals of this network are (1) to create a community of mutual trust based on an agreed ethical and juridical framework that will allow us to exchange data and (2) to understand the technologies that allow us to create a joint access domain. The reason for focusing on these two aspects are the necessity of improving the conditions for the long-term preservation of the stored unique material and the knowledge that different archives host material about the same languages or those that are spoken in neighboring communities. Researchers want to see all resources of a specific language or want to study the influences between languages without being bothered by all kinds of organizational and technical boundaries.

DAM-LR (Distributed Access Management for Language Resources) [7] is a European project where four archives serving different communities such as fieldworkers, phoneticians and computational linguists are taking practical steps to come to a joint virtual archive. All four partners have been investing substantial funds to form full-fledged language resource archives according to the above mentioned principles. The project has already worked out solutions for the essential pillars of an archive federation and is currently busy implementing them:

- (1) establishing a domain of trusted servers and services by setting up a PKI system [8] based on EUGridPMA certificates [9] (this mechanism is supported world-wide);

- (2) establishing a joint domain of metadata by making use of the IMDI metadata infrastructure [10] (due to the support for research and management shallow metadata sets as Dublin Core are not sufficient);
- (3) establishing a joint domain of Unique Resource Identifiers based on the widely used Handle System [11] where each archive manages its own URID sub-domain and therefore is free to specify the syntax of its URIDs;
- (4) establishing a distributed authentication and authorization system where the authentication is left to the home institutions of users and where Shibboleth [12] is used to exchange user credentials to allow authorization.

The federation includes a number of agreements between the partners such as

- a) agreeing on the user attributes that are exchanged when determining access rights;
- b) associating the access rights information with URIDs and thereby assuring that the owning institute defines the access rights for all copies;
- c) creating mirror sites for resolving handles;
- d) using Shibboleth for exchanging about users, but leaving the decision about the authentication system itself to the partners.

The partners identified the need to develop a resource managing component that interfaces with the other components, implements the access policies defined and an advanced access specification management component that can be used by archive managers and depositors to specify policies and access permissions. All specifications for the agreements have been made and have now to be tested in reality.

Where possible, DAM-LR is relying on components that have already shown their robustness and reliability. Shibboleth will be used although we foresee that the typical scenario where authorization is done based on user classifications such as “being a member of a student class” or “belonging to a certain staff category” will not apply to most cases in our domain. A problem may emerge at large universities where the user attributes are defined at a high university level. Departments participating in DAM-LR will not be able to convince the university boards to change the rules and also store attributes specific for the DAM-LR scenario. A simple solution can be realized when for instance LDAP is applied for authentication. A local copy with filtered information could be created and the necessary attributes could be added under local responsibility.

5. Community State

It is obvious that modern language resource archives can only tackle the above mentioned problems, since different initiatives have driven the language resource community during the last decades. Language resource specific standardization efforts have been taken by initiatives such as TEI [13], EAGLES [14] and ISLE [15]. However, only initiatives such as ISO TC37/SC4 [16] have recognized the necessity to specify generic models and schemas. It is obvious that proposals such as LMF (Lexical Markup Framework) are needed to achieve some of the goals. We also can build upon the standards developed by Unicode [17], W3C [18], ISO [19] and OAI [20] with respect to unified character encoding, the XML language to describe document structures, unified language codes, metadata harvesting protocol and many others. With respect to building federations we can build upon the knowledge and tools developed within the digital library community and Grid initiatives such as GGF [21].

6. Summary and Conclusions

The language resource domain is confronted by an enormous increase of interrelated resources that have to be managed. We foresee that this task in all its respects can only be carried out by new types of centers which we call “language resource archives”. These archives are dealing with digital material, where the rule that physical archives may not be touched (in order to preserve them for future generations) is not applicable anymore. Therefore, these new type of archives should offer services for extensions and enrichments, guaranteeing however, that the original content will not be affected.

In this paper, we have described a number of principles that these archives need to follow to offer smart and stable access services, including the primary task of long-term preservation. Currently, as far as we know no language resource archive fully adheres to these principles, but due to the previous standardization work and the experience in this field we are optimistic that we now can build such archives. In addition, these archives need, in the next few years, to form or affiliate to federations of archives. Currently, we know of two closely collaborating initiatives which are discussing and testing federation methods. Within a few years we expect to see the first full-fledged digital language resource archives to be operating within such federations. These will offer the researchers a much more integrated and accessible domain of language resources.

7. References

- [1] D. Schüller: www.mpi.nl/LAN/vol_01/lan_v01_n03.pdf

- [2] <http://www.iasa-web.org/iasa0013.htm>
- [3] <http://www.ltg.ed.ac.uk/~ht/rhodes.html>
- [4] <http://www.kb.nl/>
- [5] <http://www.elda.org/>
- [6] <http://www.delaman.org/>
- [7] <http://www.mpi.nl/dam-lr/>
- [8] <http://www.pki-page.org/>
- [9] <http://www.eugridpma.org/>
- [10] <http://www.mpi.nl/IMDI/>
- [11] <http://www.handle.net/>
- [12] <http://shibboleth.internet2.edu/>
- [13] <http://www.tei-c.org/>
- [14] <http://www.ilc.cnr.it/EAGLES96/>
- [15] <http://www.mpi.nl/ISLE/>
- [16] <http://www.tc37sc4.org/>
- [17] <http://www.unicode.org/>
- [18] <http://www.w3.org/>
- [19] <http://www.iso.org/iso/en/ISOOnline.frontpage>
- [20] <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [21] <http://www.gridforum.org/>

Appendix D: MPI LREC contribution

LAMUS – the Language Archive Management and Upload System

**Daan Broeder, Andreas Claus, Freddy Offenga, Romuald Skiba, Paul Trilsbeek,
Peter Wittenburg**

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

Abstract

LAMUS is a web-based service that allows researchers to deposit their language resources into a language resources archive. It was developed at the MPI for Psycholinguistics for stricter control of the archive coherence and consistency and allowing wider use of the archiving facilities without increasing the workload for archive and corpus managers. LAMUS is based on the use of IMDI metadata standard for language resources and offers metadata search and browsing over the archive.

1. Introduction

The language resource archive at the MPI for Psycholinguistics stores digital language resources from the institute's groups for acquisition, gesture and cognition studies and also houses the corpora of related projects such as DOBES [1] and DBD [2]. Due to newer and increasingly cheaper technologies for recording, digitization and storage the archive has now reached a staggering, at least for the domain of language studies, total of 15 TB comprised out of 150000 individual objects. This amount is ever increasing due to the 60 expeditions per year from MPI and DOBES teams that bring back an average of 30 tapes.

The archive contains a large variety of different linguistic data types, i.e., (annotated) media recordings or text sequences, lexica, series of photos, field notes, sketch grammars, ethnological notes etc. Most of the archive is comprised by digitized recordings: both audio and video and the files containing the transcriptions and analysis. Next to these, there is IMDI metadata describing the individual resources as also their mutual relationships and dependencies. The relationships between resources are embodied by embedded links in the metadata [3,4].

The institute used to be able to manage the archive with a sizeable group of corpus managers that took care of the whole process of archiving from digitizing the media tapes, moving the files into the archive in suitable linguistic determined groupings and adding the metadata (provided by the depositors). Also the corpus managers were responsible for updating existing content and maintaining specified access policies. In fact they were and partly still are the only interface between the researcher/depositor and the archive.

2. Changing the Data Ingestion Workflow

Some time ago we deliberated the possibility of a different workflow for ingesting resources into the archive, one that relies on more involvement of the depositor, using modern web-based services integrated closely with existing archive access services and procedures. There are several arguments for changing to such a system, that we call LAMUS (Language Archive Management and Upload System).

Increasing costs.

The enhanced possibilities for recording, digitization and storage also increase the workload for corpus managers. There is no balancing force against the creation of raw unanalyzed material that is stored in the archive for possible future processing and analysis. This can be worthwhile data nevertheless but some minimal description and analysis of this data should be available before accepting it into the archive.

Using Depositor Knowledge

The depositor is the best qualified person to determine the way his resources should be integrated into the archive. However he may be not the best qualified person to deal with the physical realities of the archive like file systems and setting access permission. Therefore corpus managers performed this task, but needed much interaction making it questionable if it really saved the depositor that much time.

Remote Archiving Service

In the age of the internet and web based services we see a huge potential for offering remote archiving services. Many projects are already distributed i.e. have researchers with affiliations of different universities and institutes. Using a remote archiving service they will be able to ingest their data in a central archive profiting from essential services as guaranteed backup and access.

Stricter Checks

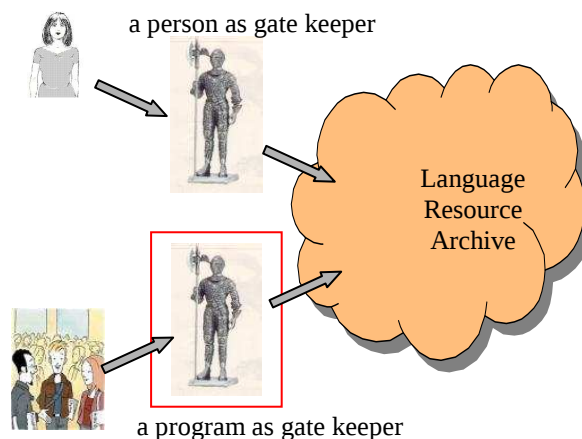


Figure 1 indicates the basic problem each language resource archive is confronted with. While until now individuals approached the archive manager as gate keeper, to take care of integrating objects into the archive, now we are confronted with a much larger group of depositors and much more data. This requires software that takes over the role of the archive manager as gatekeeper.

In the old system much depends on the knowledge of the corpus managers concerning archive policies such as what are the policies like resource naming, acceptable formats etc. At the MPI there is a big reliance on student-assistant work for corpus manager tasks, who tend to have very short-term contracts and often makes for less than perfect knowledge transfer. An automatic system that monitors the type, formats and interrelations of the ingested data can be a better gatekeeper and guarantee the archives coherence and consistency. This enormous change is indicated in figure 1.

Maximizing Deposition

According to an overview made on request by the UNESCO [5] a large amount (80%) of very important data about cultures and languages are in danger to be lost forever, if they will not be handed over to powerful enough digital archives. Of course, the MPI and the DOBES archives feel the necessity to open their gate for contributions from third persons. However, this can only be managed when the load on the archive managers will not increase, i.e., a software controlled upload option is a prerequisite to solve the huge problem of loosing data.

It may be clear that the above arguments are related only to the archive data ingestion process, they are independent from those for advocating web-based services for access and utilization of the archived data. We think the case for the last has already been proven and we won't go into that here except to describe these services as a complement to the ingestion system where needed.

3. Depositor Guided Data Ingestion

As functional requirements for LAMUS we considered the existing archive workflow for data ingestion and listed the actions that presumably can be managed by the depositor such as:

- Uploading and naming individual resources (media, annotations, information files)
- Specifying the metadata and mutual relations for and between resources .e.g. IMDI resource bundles.
- Creating relevant linguistic groupings for the data, naming and arranging the material in sub-corpora.
- Specifying the access rights and policies for the deposited material. Required functionality is the possibility to specify access for specific known groups and users as also specifying requirements for users to first sign a code of conduct before they can access the material.

- Downloading individual resources or whole sub-corpora for the purpose of updating or local analysis and uploading it to its original location in the archive.

The system would then augment the depositor actions by:

- Carrying out many checks to guarantee consistency and coherence with the archiving rules (accepted formats etc) when uploading resources.
- Carrying out typical management operations such as updating databases and indexes and creating statistics.

4. Infrastructure requirements for LAMUS

Since these upload and management services are a part of the total archive infrastructure they also have to implement a number of requirements related to infrastructure:

Universal Resource Identifiers (URIDs)

The MPI's archive has decided to introduce stable identifiers for its resources. The problems pertaining to the use of URLs are well known [6], therefore a decision was made to use the Handle System (HS) of the CNRI [7] to provide 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 publications.

Versioning.

The “stable identifier” issue from the previous point makes no sense if the resource itself is modified. Therefore, the original resource should never be deleted 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 uploaded and the depositor specifies they are to replace existing ones, LAMUS needs to first move the old resources 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 objects we have access to its older versions.

Distributed Authentication

Although the MPI archive aims at self sufficiency, we are part of different projects and organizations such as DELAMAN [8] and DAM-LR [9] that aim at cooperation at different levels. Firstly, the cooperating archives share a group of users that would like to access resources housed at different places without maintaining different user accounts. Therefore the archives should accept each others authentication of users. An accepted solution for this is the Shibboleth system [10] that will be used within DAM-LR. Secondly, the cooperating archives can host copies of each others data for safety, preservation and availability reasons.

Modularity

The MPI has offered LAMUS to be installed at other interested archive organizations. Since the needs and available resources vary considerably amongst archives, for instance not every archive is prepared to maintain a URID infra-structure, LAMUS is set up in such a fashion that such functionality is an optional addition.

5. LAMUS Functions and Workflow

LAMUS is a completely web-based service that can be used by all main-stream web-browsers. Its main functions and usual steps in the workflow are:

- 1) Allow a user to apply for an account (if none has been issued yet) by specifying his identity, affiliation, what kind of data is going to be uploaded and where the data should be linked to in the logical organization scheme. This request has to be approved by a corpus-manager, and in some cases it may be necessary to ask the advice or permission of boards.
- 2) Once this request has been accepted the user is able to create one or more workspaces where the researcher can upload resources and metadata descriptions and do all sorts of manipulations as long as the maximum allowed storage capacity is not overwritten. The user can specify relations between all uploaded components in the workspace to create a proper corpus. At any step the user can check the state of his work.
- 3) When finished for the day, the user can suspend the workspace and reconnect to it another time and continue working.

4) Once the user has finished all uploading and manipulations, he can submit a request to move the data into the archive and at that moment further checks will be carried out to guarantee the compliance with archive standards and rules.

5) When data is moved into the archive, it will also move into the domain of URID addressable objects and therefore all embedded URLs need to be replaced by URIDs. LAMUS will also take care of necessary versioning operations.

6) Relevant databases will be immediately updated so that all ingested resources are visible for everybody via the metadata browsing and search infrastructure. In our archive metadata is open, however, access to resources themselves is barred by default unless the user has specified otherwise by setting special rules for this corpus.

7) Changing the default access permissions can be done by using efficient means, i.e., the user can choose the top node of a sub-corpus and specify in one single operation that all annotations thereof should be open to the world. An access management system component is part of LAMUS and its functionality has already been described elsewhere [11].

8) LAMUS will also automatically update index files that support fast metadata and content search, although the latter is restricted to text formats for which suitable parsers are available. Content search on annotations is supported by ANNEX [12], a web-tool developed at the MPI for viewing annotation files. The upload of resources will also trigger the update of a large index that will speed up content searching.

Once the resources and metadata have been ingested in the archive they can be downloaded either individually or as a “local” corpus by special tools. The resources and metadata in the downloaded corpus keep all their interrelations by adapting all embedded links to the new situation. LAMUS allows for such “local” corpora to be uploaded again into the archive and recognizes the existing embedded links, this minimizes the construction phase in the workspace. The workflow is shown in Figure2. LAMUS is shown as a shell around the archive allowing users to create workspaces initialized with existing data from the archive (1), uploading new data into the workspace (2) and finally copying the data from the workspace into the archive (3). Figures 3 and 4 show (part of) the LAMUS user-interface.

6. Conclusions

A core LAMUS system has been operational with increasing functionality since 2005 [13]. The experiences of the users, both from the MPI and external users have been guiding the further development. Currently we are implementing the URID and versioning additions which we plan to finish this year.

The use of LAMUS is thought also to be able to increase awareness at the depositors side about the resources to be deposited. Think for instance of a Shoebox [14] lexicon that comes along with a structure file and even language files, without it the data is not complete. However the researchers is not always aware of this and in our archive we found very little shoebox files accompanied by such structure files. If possible, we see possibilities for LAMUS to guide the depositor here and explicitly demand if such structure files are also available if he uploads a shoebox file.

There are many more of these cases where LAMUS should be aware of the possible or even required existence of auxiliary files.

A necessary extension of LAMUS not described in this paper, is to make programming APIs available that allow advanced tools to directly interact with the archive without going through the phases of creating workspaces and explicitly uploading resources. For instance the lexicon tool LEXUS [15] that has its own workspace and guidance mechanism for resource creation. It needs to use LAMUS functionality directly to ingest the lexica in the archive.

To test the portability of LAMUS we recently installed it at Lund University. This was an excellent exercise to see that within half a day the complete infrastructure including some corpora from Lund University was up and running [16]. The corpus can be viewed via Internet and the researchers at Lund University can upload new resources. A training course was held to show users and archive managers how to work with LAMUS.

7. References

- [1] <http://www.mpi.nl/DOBES>
- [2] <http://www.ru.nl/dbd/start.html>
- [3] <http://www.mpi.nl/IMDI>

- [4] 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
- [5] D.Schüller: http://www.mpi.nl/LAN/vol_01/lan_v01_n03.pdf
- [6] Erickson, John. "Digital Object Identifier", In McGraw-Hill Yearbook of Science & Technology 2003.
- [7] <http://www.handle.net/>
- [8] <http://www.delaman.org/>
- [9] <http://www.mpi.nl/dam-lr>
- [10] <http://shibboleth.internet2.edu/>
- [11] A. Claus, Access Management System. *Language Archive Newsletter*, 1(1), 5
- [12] <http://www.mpi.nl/annex/>
- [13] Claus, A ,Wittenburg, P ,Broeder. D. (2005) *Language Management and Upload System*. 2nd Language Technology Conference L&T 2005, Posen.
- [14] <http://www.sil.org/computing/shoebox/>
- [15] <http://www.mpi.nl/lexus>
- [16] http://dam-lr.sol.lu.se/ds/imdi_browser/

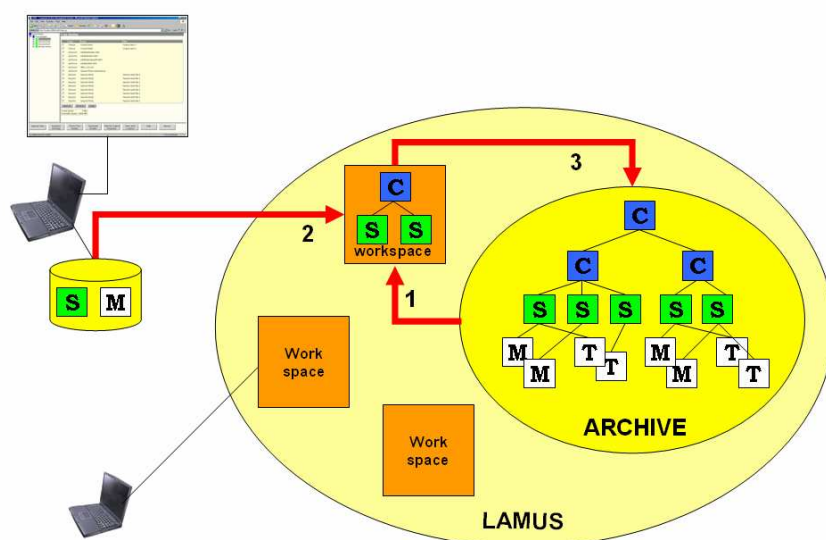


Figure 2 indicates LAMUS the basic workflow. New resources can be uploaded from a notebook or another archive into the workspace and from there into the archive. A user can also copy archive resources into the workspace for further processing and then upload them again as new versions.

The icons stand for [M] media and [T] textual resources, corpus metadata [C] describes the linguistic groupings and resource metadata [S] describes resources and their interrelations.

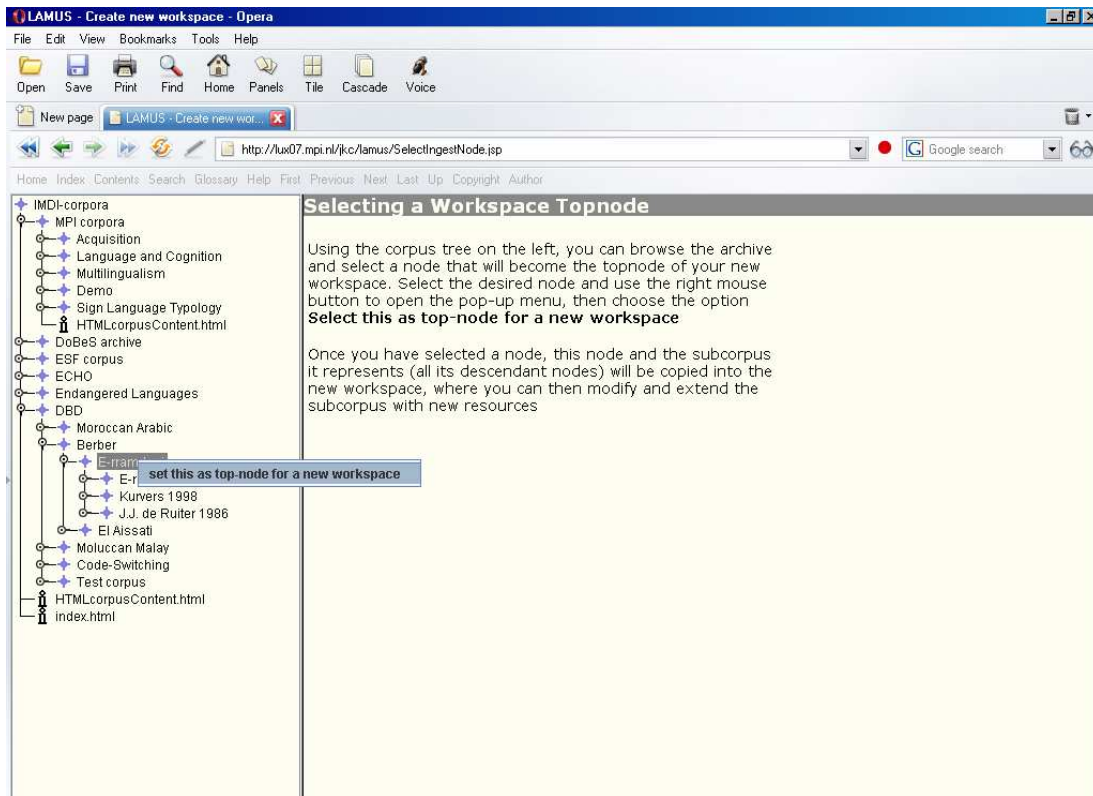


Figure 3 how part of the archive is selected to initialize a new workspace.

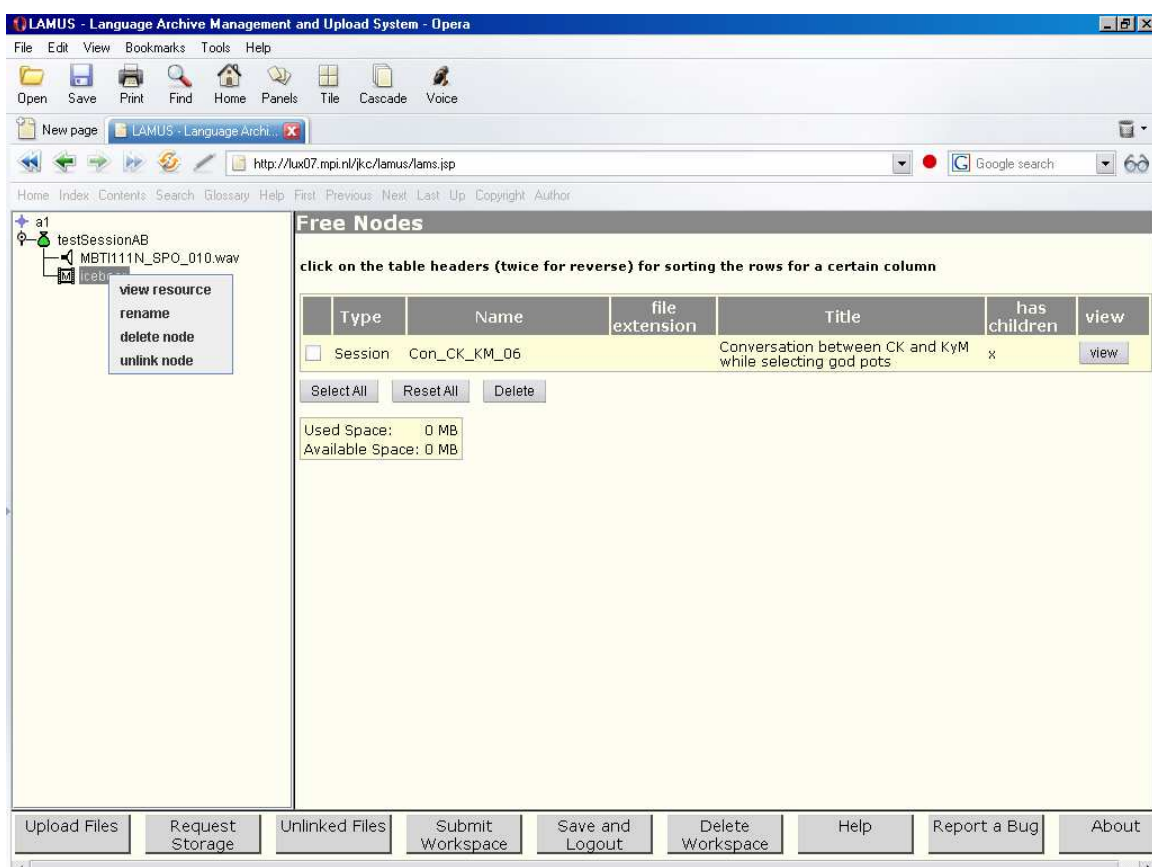


Figure 4 shows a view at a LAMUS workspace (left) and at the list of uploaded files (right).

Appendix E: MPI LREC contribution

Comparison of Resource Discovery Methods

Alex Klassmann, Freddy Offenga, Daan Broeder, Romuald Skiba, Peter Wittenburg

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

Abstract

It is an ongoing debate whether categorical systems created by some experts are an appropriate way to help users finding useful resources in the internet. However for the much more restricted domain of language documentation such a category system might still prove reasonable if not indispensable. This article gives an overview over the particular IMDI category set and presents a rough evaluation of its practical use at the Max-Planck-Institute Nijmegen.

1. Introduction

The raw material for linguists are samples of a particular language. These may range from pieces of parchment til recordings of TV broadcast. Although there exist guidelines for the metadata description and annotation of linguistic resources (IMDI [1], DC/OLAC [2], TEI [3], EAGLES [4], specialized data bases), no standard is universally accepted and probably can't be since researchers will focus on different aspects and invent new theories and ideas. The amount of collected and electronically available resources has exploded over recent years and poses the problem of organization/management and (re-)discovery of the data. In this paper we will present the approach the MPI for Psycholinguistics has chosen with respect to the metadata description, will elaborate on a number of different location methods and finally will discuss some critical points. The first paragraph will give a short overview over the IMDI metadata scheme. Then their practical application i.e. the tools which allow the user to handle this metadata set will be presented. A rough evaluation of the quality of the at present available metadata follows. Then an alternative to formal categorization will be presented, namely free „tagging“, which is currently lively discussed with respect to internet search engines. Its applicability to the field of linguistics will be questioned and some preliminary conclusions drawn.

2. IMDI Metadata

The IMDI (ISLE MetaData Initiative) scheme was developed during 2001-2003 by a broad network of linguists from different sub-disciplines such as field linguistics, phonetics, multimodality research and corpus linguistics. Its purpose is to give a solid, precise and extensible framework for the organization, bundling and retrieval of in principle any kind of digital linguistic resources, in particular annotated media streams and text sequences making up by far the largest percentage of current resources in language resource archives.

Typically primary language documents like audio or video files are accompanied by one or more text files, containing a transcription, translations and annotations at other linguistic levels (morphosyntax, semantic, etc) of the former and seen in the IMDI framework as resources themselves. An IMDI-session contains a detailed meta description of those tightly connected resources, and could therefore be named equivalently as metadata about a 'resource bundle'. The IMDI-schema describes in addition how those sessions can be grouped together into corpora and sub-corpora. Although corpus organization is relevant for management and browsing, it is not of relevance in this paper, i.e., for more details we refer to other IMDI documentation [5,6].

An IMDI-session can be best thought of as a form with roughly 150 hierarchically ordered entries, which concern e.g. information about

- the event (recording location, date, etc),
- the languages involved,
- the speaker(s),
- the type and nature of speech,
- technical information about the resources and
- access rights.

For most fields one or more values can be selected, but there are also so-called descriptive fields for the input of free text. Furthermore there is the possibility for every user to add arbitrary key-value-pairs which can be

interpreted as a personal or project-specific extension of the schema. In order to facilitate the procedure of filling in the metadata, a special professional editor has been build at the Institute.

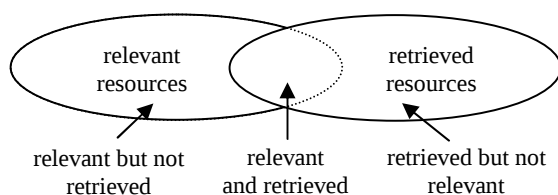
A single field, the „bundle name“-field is obligatory, yet users are urged to fill in all others, too. Unfortunately they tend to avoid this time-consuming work oriented to a re-usage by others and fields stay empty or have a default setting. Although everyone agrees that filling in metadata is very important in many respects, in particular since the knowledge about the content may be lost within shortest time, the amount of time spent on this aspect in the whole resource management life cycle is still too little.

3. Methodological Issues

One important question for the usage of archives – traditional as well as modern – with an extremely growing amount of resources is the possibility for the user to locate useful resources. As described the MPI uses the structured IMDI set to describe resources which therefore lends itself to carry out queries. Metadata includes added value with respect to the resources themselves, therefore it is data that cannot be missed. A recording may include an interview with a person having certain characteristics such as age, sex, education etc. Only in rare situations the recording will contain this information explicitly – it is the metadata description that will allow the interested user to make a comparison between male and female language use for example. Many other examples of this added value can be given.

Although we will have very different user groups ranging from researchers, teachers, students, journalists to the speakers themselves. All have different types of queries and all asking different types of interfaces. Nevertheless, we can make a few general statements on what a typical search method should optimize.

Literature defines two terms, “precision” and “recall”, as measures for the success of a query. With “precision” the proportion of hits that are relevant compared to the irrelevant hits is meant. A higher amount of “noisy” results would therefore reduce the precision rate. With “recall” the proportion of relevant hits that were found compared to the not found relevant hits is meant. A query method that would not find very much of the relevant resources a user is looking for obviously would be not successful. The following drawing taken from G. Simons [7] is very useful to indicate the relation between the two terms.



Another important point in searching is of course the question of how to rank the hits. The precision could be very low, i.e., the number of irrelevant hits could be high, but if the relevant resources would be presented at the top of the list the user probably wouldn't bother. In this paper we will not discuss the ranking aspect.

4. The MPI Archive

The Max-Planck-Institute Nijmegen houses a digital archive with a large variety of different language corpora, all categorized with the IMDI metadata set. The archive encompasses ca. individual 45.000 IMDI-sessions describing about 150.000 resources.

Infrastructure and tools have been designed to offer to the user several options to search for a specific IMDI-described resource. Since metadata is open per definition, all descriptions are accessible via the web; cf. http://corpus1.mpi.nl/ds/imdi_browser):

1) Browsing in linked resources. This is similar to clicking through a local file system with the difference that the hierarchy of corpus structures is much more stable. The approach is aimed at users familiar with or quickly able to grasp the underlying logical organization. Bookmarks help to make this process more efficient.

2) Structured search within the whole archive as well as within a selected part of it. Every IMDI-element can be addressed individually and the search for different elements can be combined into one query. Queries like "Give me all video files that show a female Wichita speaker older then 60 years" can be formulated and a high precision, i.e., a low number of irrelevant hits, can be expected. Yet, the user has to know the terminology used by the IMDI schema in order to achieve a high recall, i.e., get a high percentage of the resources having looked for as hits. Furthermore, search is restricted to elements with closed or open vocabularies and does not cover elements with free text.

3) Unstructured search over the whole or part of the archive. The user 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. It is possible to formulate logical combinations of expressions and even "fuzzy terms" (for an overview of the possibilities cf. [8]). The recall with this method can be expected to be higher compared with structured search, however, the precision will be poor, i.e., much more irrelevant hits can be expected.

4) An extension of unstructured search is to provide the metadata descriptions to web search engines like Google with their advanced information retrieval techniques. However search cannot be restrained to a specific corpus, not to mention parts of it, and results will include a huge amount of unwanted hits from the whole internet. An additional term such as „IMDI“ or „MPI“ improves the precision significantly, but still yields unsatisfactory results.¹

5) All IMDI records were transmitted to the OLAC service provider (DC [9]). OLAC offers a structured search possibility, but limits itself to the elements of DC and a few additional ones such as the language a resource is in. Currently, the service is not working well, since the OLAC service provider does not accept too many records, i.e., they expect the data provider to just deliver one metadata record for a sub-corpus. For the MPI it is in many cases difficult to determine what exactly a sub-corpus is. With respect to precision and recall we expect similar results as with structured search, as long as the restricted set of elements is sufficient. An advantage of using OLAC, however, is that other archives will contribute to OLAC, too.

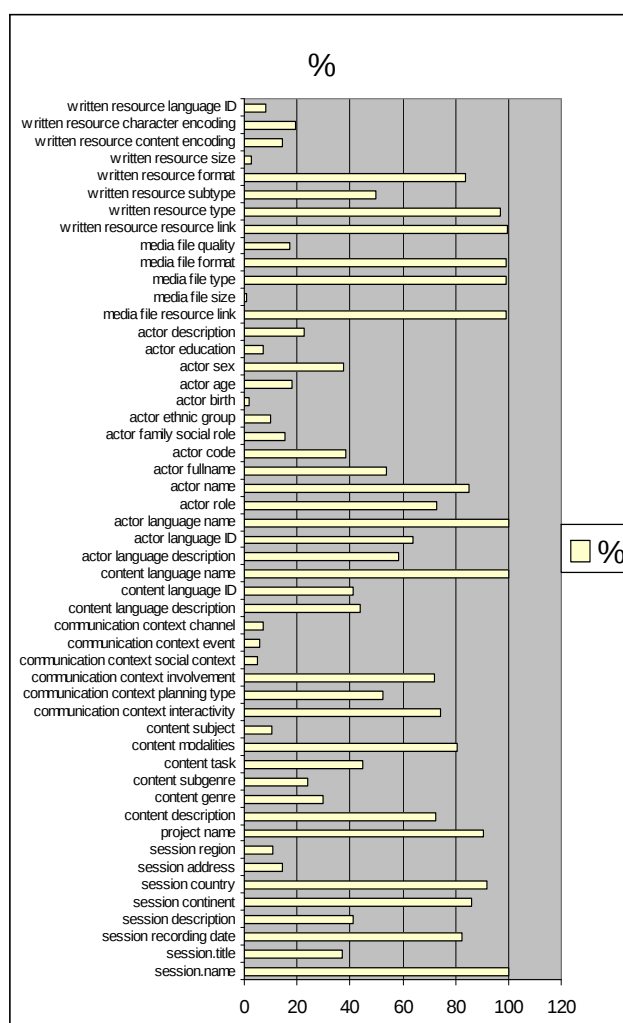
6) Geographically orientated browsing. Since many languages in the archive are related to diverse and less known regions all over the world, a geographical browsing makes sense, too. The visualization tool Google-Earth [10] is used for this purpose, where the user can look for spots on the physical map of the Earth that point to IMDI-files. Of course, this method yields an enormous high precision and recall if only the geographic location is the discovery criterion. Since this approach is of less theoretical interest, we will not elaborate on this option.

We should not forget to mention that in general researchers want to combine metadata search/browsing with searching on the content as it is possible now for example with ANNEX [11]. Typical questions such as “give me all instances where a 4 year old female speaker is using a certain morpho-syntactic construction” can only be addressed when a combined structured search is performed. But we also understand that such questions will only be addressed by the “very well informed” user who knows exactly the terminology that is used. All other search options will not lead to useful results. In this paper we will not include the content search option, but discuss metadata search options in general.

5. Evaluation

In order to have significant variance in the data, an evaluation of the metadata was done on a subset of the resources in the archive, where metadata was filled in manually and by different users, i.e., the Dutch Spoken Corpus, for example, was not included.

The table below gives an impression of how often fields are actually filled in (e.g. not empty and not default values like „unknown“ or „unspecified“). These statistics were created on 23.710 resource bundles. As can be seen the sets are far from being complete. On the other hand, every field of the scheme (including those not



¹ When searching for example for real resources for the TEOP language a Google search with “teop” as query string yields 17.600 hits with lots of unusable hits. A query string “imdi teop” only yields 683 hits and more important the entry for the Teop corpus is amongst the first five. However, users suffer from the same deficit: how should they know which string to use to achieve an acceptable precision and recall.

shown in the table) has been used in some sessions, so that it seems that no field in the schema is obsolete. These statistics give a baseline idea of what can be expected.

Since there is still not sufficient experience at the institute with actually performed metadata searches, it is not yet possible to carry out a full-fledged statistical evaluation based on empirical data. Instead, test queries which might be of relevance for researchers were formulated and executed. It was then checked whether the hits were accurate.

So, e.g. in Second Language Learning Research the influence of age on the acquisition of language is examined and it is assumed that there is a critical period in childhood for the development of certain skills such as learning grammar constructions. In order to find resources one would like to formulate a query like „Give me all resources for a given (not-mother-)language for speakers aged between 4 and 16 years“. Since the development between boys and girls may differ one even could refine the query by an appropriate qualifier.

Using the IMDI structured search the following query “Language=Dutch, Actor.Language.Mothertongue=false, Actor-Age<16 and >4” yields 203 hits. An additional selection on “Actor-sex = Male” results in 119 hits and one with “Actor-sex = Female” in 83 hits. A full-text search with a query “Dutch AND second AND language AND (15 OR ... OR 5)” results in 488 hits and may be still useful, too.

Categorization with respect to age and sex as well as technical categories like the file format are rather uncontested and not prone to subjective interpretation. This is different with respect to the descriptive elements concerning the content. Here the difficulty can be seen at the many corrections the initial IMDI set experienced and the user is merely offered a list of given values, but can type in others (“open vocabulary”).

The vocabulary for the element „Content-Genre“ e.g. encompasses 13 items („discourse“, „poetry“ etc.), two of them never have been used („Popular fiction“, „Newspaper article“) and another 15 values have been added by users. Concerning the element „Content-SubGenre“ the situation is similar: no offered type of drama has been used and (fortunately!) no resource was classified as „Unintelligible Speech“. Some 30 items were added, ranging from broad terms like „Speech“ to very specific ones. This poses the question if such a categorization in advance by a group of „experts“ is the right approach for data organization.

6. Free Tagging

In this paragraph we will discuss free user „tagging“ as opposed to categorization based on an a priori defined categorization schemes.

With respect to searches in the internet the early stage approach from Yahoo to perform search along given categories has been abandoned in favour of key word search as known by Google. Yet simple string matching in documents is not very precise and doesn't work at all for media files. Currently an alternative to in-advance categorization might be 'user tagging' as it is promoted most outstandingly by Shirky [12]. He refers to a service [13] that offers users to store bookmarks of web-resources and make those bookmarks available for the public. So each user who wants to remember an URL of interest can describe it with an arbitrary set of key words. Of course, each user has his own view of the resource and the description may be inaccurate or erroneous, but the assumption is, that if there are a lot of users describing the same URL, the statistics will end up establishing a widely shared set of key terms. This kind of „categorization afterwards“ lacks genuinely any hierarchy and results more in a kind of semantic net or „topic map“.

7. Discussion

There are a number of reasons why the idea of “free tagging” will not be applicable for the domain of language resources:

- The idea of „free tagging“ relies on the voluntary work of many and presupposes that the resource in question is interpretable by everybody. This is certainly not the case in the field of linguistic data, where often only the producer of the resource is able to describe it adequately.
- It is the researcher who has the deep knowledge about the construction of a corpus and about the reasons to have chosen a certain approach. This knowledge has to be stored somewhere and it's the metadata where it is stored.
- At least the linguistic users can rely on the a priori defined categorizations, since linguistic terminology has stabilized to a large extent during the last decades.

So, tagging of the content of linguistics resources would have primarily to be done by the creator like with the rest of the metadata. On one side, the „open vocabularies“ offered currently by IMDI incite some users to

slightly misuse them for an imitation of „free tagging“ e.g. if they add an overspecialized item. On the other hand “free tagging” could be an option for other “experts” to enrich the data and therefore to increase the precision and recall.

A solution and kind of promise between the two strategies may be to make every new entry „public“, e.g. adding it to the list of offered vocabulary automatically. This would benefit those who fill in the data as well as those who are querying it. Furthermore, it would inhibit users to add too specific terms by a kind of „social pressure“.

8. Conclusion

The Max-Planck-Institute Nijmegen offers several kinds of querying and browsing approaches corresponding to different user interests. The IMDI categorization scheme allows in principle for very detailed search and therefore has the potential for a high precision and high recall compared to all sorts of free text searches.

However, the IMDI forms are generally not completely filled in as was indicated in the table and even linguistic users do not fully share the same terminology. This will deteriorate the success of the searches in terms of precision and recall. Since free-text field also bear relevant information in many cases, even some linguists will prefer nevertheless a free-text search on the metadata first.

9. References

- [1] <http://www.mpi.nl/IMDI>
- [2] <http://www.language-archives.org/>
- [3] <http://www.tei-c.org/>
- [4] <http://www.ilc.cnr.it/EAGLES96/>
- [5] 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
- [6] Broeder, D., Wittenburg, P., Crasborn, O. (2004). *Using Profiles for IMDI metadata creation*. In X. Fatima Ferreira et. al. (Eds), *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: European Language Resource Association. pp1317-1320
- [7] Aristar-Dry, H., Simons, G. (2006). *E-Meld: openness, ontologies and interoperability*. DGFS Annual Meeting on Language Documentation and Description – Working Group 6. University of Bielefeld
- [8] <http://lucene.apache.org/java/docs/queryparsersyntax.html>
- [9] <http://dublincore.org/>
- [10] <http://earth.google.com/>
- [11] <http://www.mpi.nl/annex>
- [12] Shirky, Clay (2005): www.shirky.com/writings/ontology_overrated.html
- [13] <http://del.icio.us>