Deliverable 9.1
Specification 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: D9.1

Authors: MPI

Responsible: MPI

Date: 1.3.2006

Content

1 Introduction

The goal of this report is to present the specifications for the core components of DAM-LR based on the results of the prototype implementation. All major agreements for the pillars of the project were carefully discussed with all partners and documented in the definition report (D8.1). Some specific components need to be developed for the complete distributed solution. Details will be added to the definition report as they become available.

2 PKI System

The foundation of all distributed services are trusted servers and services. The EUGridPMA is the European authority that is accepted to establish requirements and best practices for grid identity providers to enable a common trust domain applicable to authentication of end-entities in inter-organizational access to distributed resources. As its main activity the EUGridPMA coordinates a Public Key Infrastructure (PKI) for use with Grid authentication middleware. To support this it maintains the TACAR (TERENA Academic CA[1] Repository) repository which is a trusted repository which contains verified root-CA certificates and which can be entered into local lists.
For DAM-LR this is the way to go, since it includes the certificates from

- the German DFN - the MPI is RA within the DFN domain
- the DutchGrid/NIKHEF - the INL should become RA within that domain
- the NorduGrid/SwUPKI – the Lund university should become RA within that domain
- UK eScience – the SOAS should become RA within that domain

The MPI already started the procedure to become RA which means that it can request certificates for servers and services in the DFN domain. It is suggested that the other partners also start this formal procedure if it is not already done by their university bodies.

3 Unique Identifiers

Persistent identification of resources is a major point to create a distributed archive. The following issues were discussed and agreed by all partners:

- For DAM-LR the Handle System will be taken as its basis for operating with unique resource identifiers, i.e. a handle consists of a prefix given by the CNRI[2] and a postfix to be specified by the handle authority.
- Every partner is a handle authority, i.e. every partner can decide himself about the syntax of a its handles. This requires, however, that handle requests crossing the local boundaries have to be resolved by the global handle resolving service. Caching could be used to increase performance.
- Every partner has full control about his Handle database, i.e. no one else will get the permission to change entries except via clearly defined services in the case of modifications of paths for copied data.
- Every partner therefore has to install and maintain the Handle System on a server and has to take care that its database will be maintained properly.
- For redundancy reasons the MPI will host mirrors for all partner services, i.e. in case of server problems the URIDs could still be resolved.
- There is a recommendation to not use semantics within the postfixes, but in fact every partner is free in his decisions.

The Handle System has already been tested by the MPI and seems to fulfill all requirements with respect to performance, security and manageability. It should be mentioned here, that MPI will build tools in a way that they can operate with URIDs and without.

4 Authentification

With respect to the way authentication is done in a distributed scenario a number of facts will guide our decisions:

[1] CA = Certificate Authority; RA = Registration Authority
[2] The Handle System created by CNRI is a widely used system so that we can expect reliable services in the future.

- Due to national and European law we are not allowed to distribute sensitive information such as passwords and we need user acceptance to exchange other data.
- It is general knowledge that centralized user administration across large institutions is not feasible.
- Since authentication will be just one module in a complex distributed access management system one has to rely on widely agreed standards as much as possible to save time. On this background the choice for Open LDAP[3] as the basis for local authentication is recommended.
- It is possible partner institutions/departments do not control their user administration, i.e. they have to start a discussion process of how to best create a joint domain.
- It may be necessary (see below) to have the possibility of one integrated search domain, i.e. it should be possible to propagate some open attributes of users to a trusted higher node such as it is possible with LDAP.

Therefore, the MPI will step over to Open LDAP for authentication for its own user management, which will include internal and external users. Internal users are those who have a formal contract with the MPI, external users are those who want to have access to resources stored in the archive, but don't have a formal affiliation. The DAM-LR core solution will rely on LDAP, all partners that will choose for another authentication system have to develop appropriate gateway software.

In a distributed domain the partners in a federation have to exchange user information that is sufficient to grant access to resources. It seems to be a broad experience that it is wise to agree on a minimal set of such information to limit the administrational effort and to keep the system as simple as possible. A number of exchangeable credentials were discussed such as

• first name	first name of the person which will normally be used
• last name	first name of the person which will normally be used
• affiliation	name of institution they have a contract with
• hosting institute	hosting institution for the case that the person is an external user (the hosting institution can be set by default to the address of one of the partners in a federation)
• email address	email address of the user
• status	status of a user in the institution, for externals the state can be such as guest, research fellow, collaborator
• class+	the user could be member of one or more groups such as being student of a certain class or a member of a certain tribe; there could be several groups the user is belonging to
• userID	a unique string within the federation space with the help of which everyone must be identified (it seems that this ID is not necessary per se, since name and affiliation could be sufficient, but experience tells us that it is always good to have a unique identifier in addition)

These attributes seem to be the minimal set and are largely overlapping with the specifications from EDU-Person and RFC 2798[4], others such as a host introducing an external guest as a collaborator or so may be necessary but are not yet fully identified. The MPI certainly will store internally aspects such as department, start of employment, end of employment and behavior flag. In DAM-LR we have to decide whether we will speak about accounts that are valid for a limited period of time only and whether this limitation is associated with specific requests and/or with the account itself. Both seems to be appropriate. The information about the duration of a request certainly would have to be stored together with the other request information.

The behavior flag is relevant for the MPI to indicate persons who severely misbehaved. If the flag is set all access will be ignored. We have to have such possibility to memorize such form of misbehavior. Of course, we cannot prevent completely that the same person will register again under

[3] LDAP is basically a specialized interface for database information that is typical for example for user identity information. It comes with many ready-built-modules and it is widely used in the academic world.

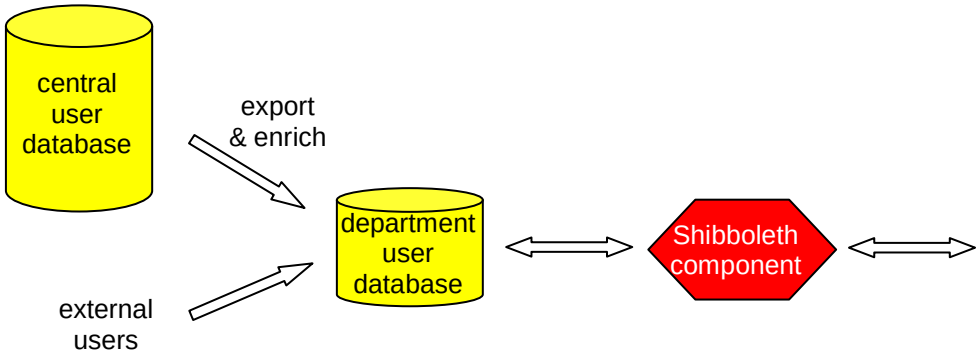
[4] For RFC 2798 there is an existing LDAP schema that could be re-used.

another name. However, when we would apply the host concept it would become difficult to sail under other names. This issue is tricky and has to be discussed.

For departments that are part of large institutions such as linguistics department at Lund university it may be two problems:

- it could be difficult to be home institution for external users, the university computer centre may refuse to accept them in their central user database
- it could be difficult to add attributes in the central user database that are required within a federation

For these cases LDAP offers a simple solution which is sketched in the following drawing:



LDAP comes with functionality that could help to implement such a scenario easily.

LDAP allows to set rights such that only certain attributes can be exported.

5 Authorization

The aspects that have to do with authorization in a distributed scenario are the most complex ones. Therefore we will split the discussion in 6 aspects.

5.1 General Aspects

The basic goals we want to achieve in DAM-LR are the following:

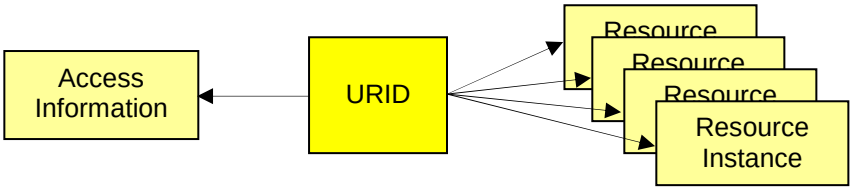
- single identity — to be achieved by distributed authentication and accepting attributes
- single sign-on — once the user is identified he/she has transparent access to all resources he/she is permitted to access in all archives
- one basket idea — the user must be able to see his/her set of accessible resources as his/her temporary working archive
- replication option — the archives must accept each other in so far that they exchange data about resources and resources themselves

One of the basic agreements in DAM-LR is that access handling is done by the originating institution. Since for each resource independent of the number of instances there will be only one URID[5], it seems to be a direct conclusion that

- the URID record is maintained at the site of the originating institution
- all access rights information is associated with this URID entry.

The URID is the incarnation of the resource. It has pointers to all instances that can be stored on different servers and it knows about the access information set for the resource which is valid for all instances.

[5] There may be resources that do not have a URID for whatever reasons, i.e. certain tools will have to work both on URIDs and URLs.



First, we have to address the question what the typical usage scenario of our archives will be. Many distributed usage scenarios that are discussed currently have the characteristic that a whole group of users will want to access resources based on the fact that they are formal part of such a group:

- all university staff members want to access all e-journals of a certain publisher
- all students of a certain class want to access certain recommended teaching material
- etc

In all these cases the users share a formal group assignment which is also part of their user entry such as being staff member or being student of a class etc. In our usage scenario we will have these cases as well, but in general we will have individuals who want to access the resources:

- individual researchers who want to analyze specific language phenomena
- students who want to write their master thesis or their PhD
- journalists who want to elaborate on a certain language family
- etc

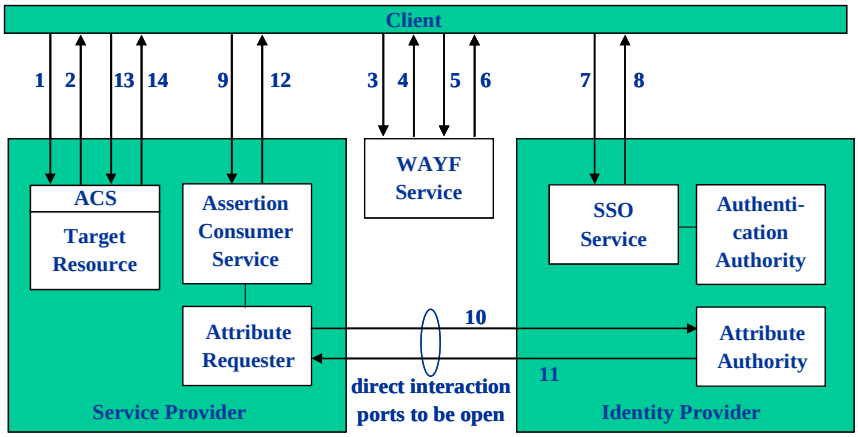
In all these cases it is not a single group marker that will give access permissions, but the individual user ID. Consequently, at the authorization side much more work has to be done to enter access permissions of the users and this sight has to know about the registered users. This has to be considered when designing software, since the administrational load can become intractable.

Another difference to the typical Shib scenario (see below) is given by the fact that users will partly request access not to just one resource but to several (search across corpora, access to annotated media files etc). This also has to be taken into account.

5.2 Shibboleth

Shibboleth is a software product that was designed to primarily facilitate distributed authorization in a scenario where groups need access and where group marks are exchanged. It was designed to help in the access scenario dominated by groups. Nevertheless, we currently believe that Shibboleth is the best component around to exchange user information in a secure way and it is increasingly often accepted by universities etc in different countries, i.e. there is a broad user community and institutions will increasingly often accept Shibboleth for the kind of trusted operations as required in distributed scenarios. One of the major advantages for us is that Shib puts responsibility for authentication at the home institute.

Let us therefore first introduce Shibboleth briefly (for details we refer to the Shibboleth documents). The following figure indicates the different Shibboleth components (as described in older documents). The essence is that the resource provider that has to handle an access request has to ask the identity provider whether the person is known and what his/her attributes are, i.e. Shibboleth has an interacting role between the most important components which are the authentication mechanism and the resource manager that finally delivers the data. For the authentication it is known for example that Shibboleth can interact with LDAP services, therefore the choice for LDAP as authentification system makes sense. With respect to the resource manager it seems that new software has to be developed since known software such as Apache are not supported. In addition, Shib expects a web-browser to request access to a single resource. In the DAM-LR scenario we also can expect applications such as content search that will request access to a number of resources.



- ACS
  Access Control Service
- WAYF
  Where are you from service
- SSO
  Single Sign-On

1 Get Resource
2 Redirect (302)
3 Get Form
4 Send Form (200)
5 Submit Form
6 Send Cookie and redirect (302)
7 Request Authentication
8 Authentication Response
9 Send an Assertion Profile
10 Request Attributes
11 Send Attributes
12 Redirect with attributes
13 Send attributes for check
14 Provide Resource

All interaction is done by creating profiles including SAML assertions.

When analyzing the information flow with respect to repeated requests a few options seem to be possible:

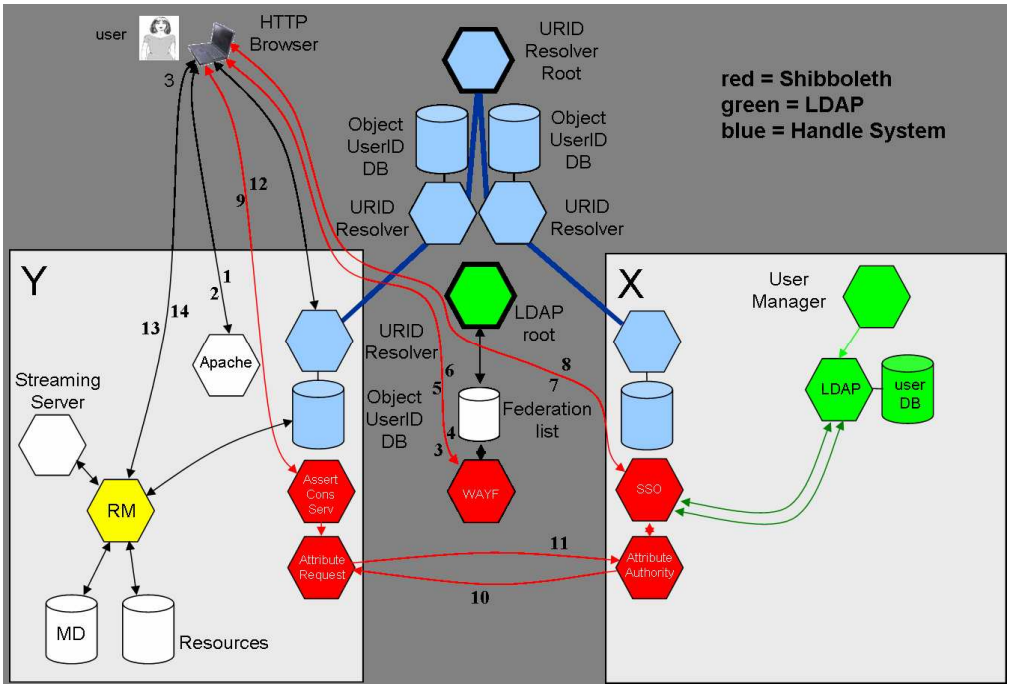
- The profile finally can contain all necessary information about a user such as user attributes, and session number. When the user wants to access another resource (1) all information is available at the client, i.e. the client could immediately step over to (13). The ACS module could directly check whether the user is allowed to access the resources and in case of matching directly deliver (14).
- Another, but less efficient option would be to just step over from (2) to (9) since the identity has been checked already.

It should be discussed in detail what the best solution will be for us.

5.3 Typical Access Scenario

The following figure indicates a typical DAM-LR flow of information.

A typical user could interact with a metadata browser, navigate in the (open) metadata domain, find a suitable resource and addresses a request to the Apache server. All interaction will include URID resolution requests. Due to configuration entries the Apache server knows that the requested resource is protected and issues a redirect (2) to initialize authentication. The WAYF service is used to find out what the home is of the user (3-6). The Single Sign-On service is contacted to let the user authenticate him/herself (7). After having interacted with the LDAP service an assertion profile is send back (8) which is then redirected to the Assertion Consumer at the service provider side (9). In the DAM-LR scenario the Attribute Requester will be contacted to ask for all open user credentials which is done by interacting directly with the Attribute Authority (10). By contacting LDAP the attributes are extracted and returned (11). The assertion Consumer returns a new profile which is then redirected to the Resource Manager (13). The resource manager will check whether the rights are ok by interacting with the Object-User Database and finally deliver the requested data.



From this figure it is obvious that the RM has to be developed. It has to interact with the following components in a safe and reliable way:

- receive an appropriate profile with attribute assertions via redirect methods from Shibboleth
- compare the attributes with the record that is stored in the access record associated with the URID (the original URID must be part of the request)
- interact when necessary with a streaming server
- deliver the data to the client

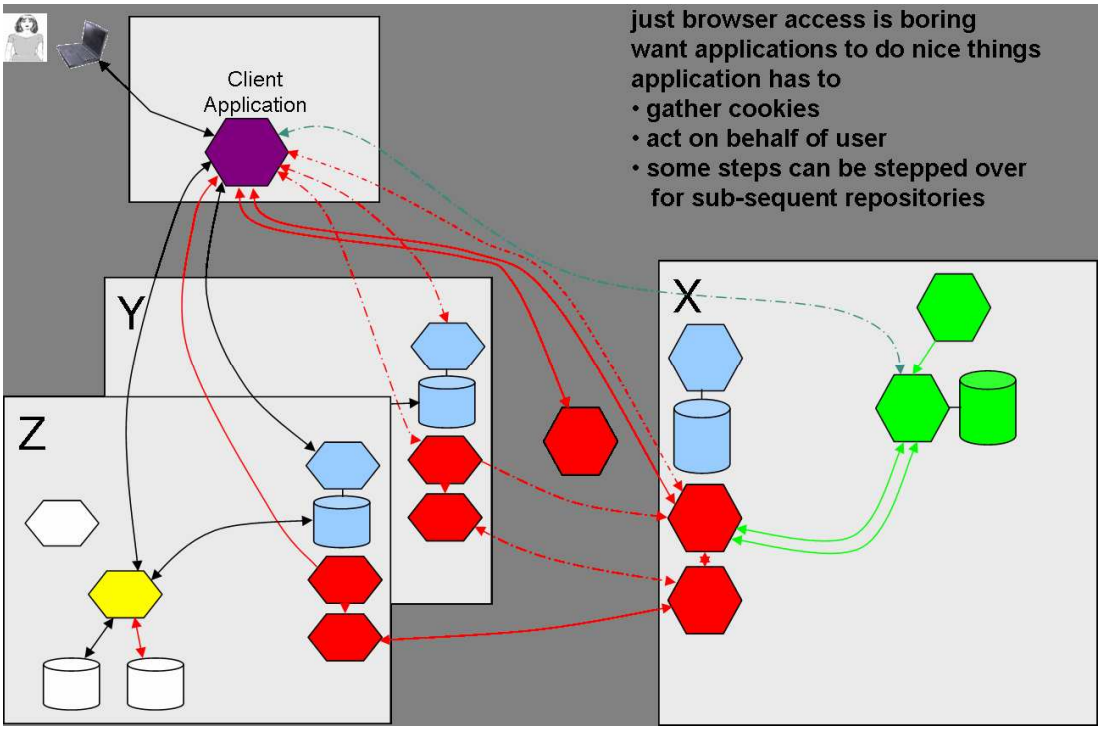
We should look for existing solutions rather than do all the programming ourselves.

5.4 Application Access

Often the users will want to access the data via web applications such as ANNEX or LEXUS to operate on complex data types and multiple resources eventually from different sites. As an example a search may be wanted that includes metadata and content from a basket of resources coming from all partner archives. This scenario is depicted in the following figure where the essential components are shown in a reduced way.

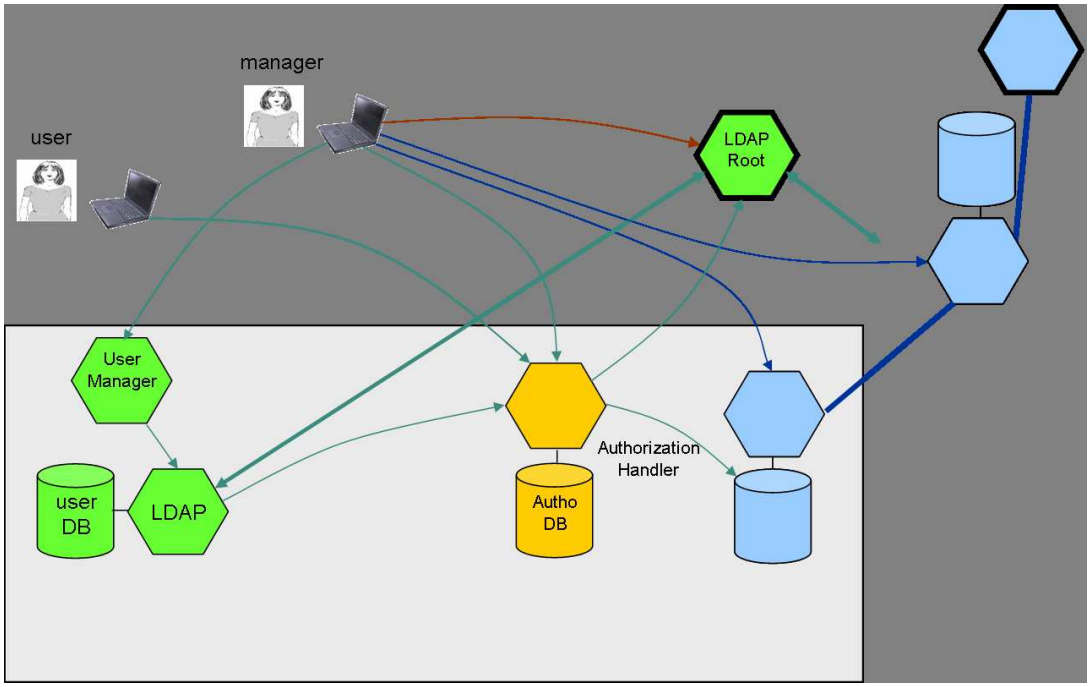
The application has now to act on behalf of the user if different sites are involved. The user name and password will be requested by every site that has to be contacted, but the user should type it in just once. To access several resources from one site the application needs to store the appropriate cookie to support access. Therefore, only for the very first interaction the user should have to type in user name and password. The application has to store this information and present it again during the next request. Yet we don't know exactly what the exact content of the cookies is Shibboleth generates, i.e. implementations details have to be sorted out.

The applications have to be extended to support such scenarios. In a basket like situation it may make sense to do the authentification for all included resources at the beginning to assure that any subsequent operation can be executed smoothly. As already indicated it has to be sorted out what the best option is in terms of efficiency.



5.5 Management Scenario

DAM-LR has to provide a feasible management framework. In the following picture some essential components are indicated.



New User

A user may want to fill in a form to get registered at an institution. In this case the manager will check all specifications and in case of external users ask for a host who can make a positive statement about the person. With all information available a new record will be generated into the local LDAP system. The record has to contain all attributes as agreed in DAM-LR. For modifications of user

records similar steps have to be taken. Of course, we have to distinguish between users from the institutions and those who are accepted as guests.

The LDAP systems of the partners can be linked so that a joint domain can be generated. Other architectural solutions are possible. It has to be decided later which of them are most efficient to implement and to maintain.

New Resource
For entering a new resource a new record has to be created in the URID database. At MPI this will be done by LAMUS which is the resource ingest software, i.e. the manager only has to control the entries. When the physical paths are changing a mover/copier has to be used to modify the record content.

User Resource Request
At the beginning of each access activity we can assume that a user will fill in a request form with a request to access a certain resource. The form will probably ask the user to enter all relevant attributes and the resource he/she is interested in. The manager receives this information and has now to find out which user the requester exactly is. He will do a search via the centralized LDAP root or via another mechnism to find out where the person is registered and whether all specifications are correct. For this purpose we will need a joint domain that contains all relevant information that may be exchanged from all sites. The manager may want to take another action – namely sending an email to the depositor of the resource – and ask for comments. In case that everything is ok the manager will create an entry in the Authorization DB for the requested resources. We should add here that resource requests could also mean that a user asks to get access to a whole sub-corpus or only to the lexica in a certain corpus etc, i.e. the Authorization DB contains commands on a high level.

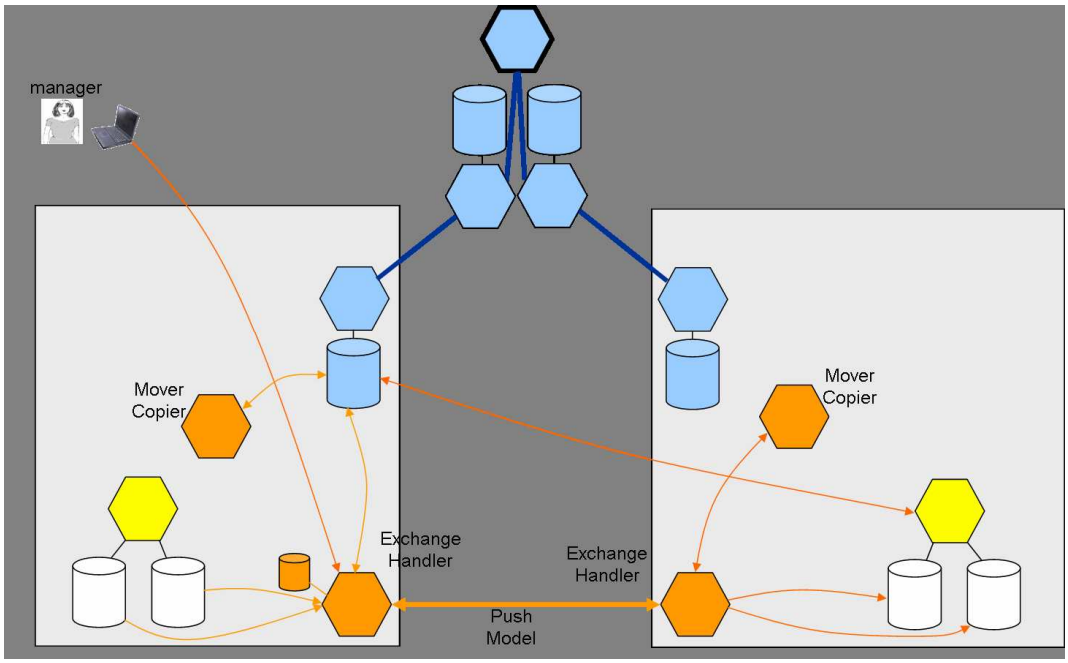
The Authorization DB also contains per sub-corpus specifications about processes such as whether the depositor has to be asked first, whether the person has to sign a declaration etc. When the user has fulfilled all required steps the general command will be transferred into corresponding formal URID access records[6] by an automatic process running at regular intervals. When this extension has been done the user can finally access the resource(s). It is up to the repository to exactly define the steps and the way management is done.

5.6 Data Moving Scenario

System Managers for example want to move and/or copy resources. Two scenarios have to be distinguished: local and remote changes. A mover/copier component has to be developed that contacts for all modifications of physical paths the URID database and modifies the entry to prevent dead links.

In the case that data will be exchanged between repositories[7] with the intention that the resources are accessible via alternative paths additional components have to be involved. First, we need a pair of trusted exchange handlers that take care that complete corpora including the structure of the data and the metadata are copied to the other site. The exchange and synchronization of data will require some form of protocol that has to be worked out, but that will not be subject of this document. Of course, the exchange handler will make appropriate entries in the URID database so that the URID resolver can offer two different physical paths after having copied the data.
Also at the mirror site the system managers will move or copy data at certain moments. We have to assure that the URID entry (there is only one per resource and this is maintained at the originating site) will be updated. Since one of the basic agreements is that URID databases may only be managed by local managers or locally controlled services, we have to provide a service (whether re-use exchange handler or separate ones) so that the remote mover/copier can interact with the remote instance and that the protocol supports this kind of modification information. The local exchange handler instance then will lead to modifications in the URID database.

[6] The exact nature of the requirements for the Authorization DB has to be discussed. At the MPI this system already is operating and has shown its robustness and administrative efficiency.
[7] In DAM-LR this is not a requirement, nevertheless, it makes sense to consider this option and its effects on all operations. Finally, it is a goal within DELAMAN to exchange data and to make it available via different channels while keeping the ownership and all access information the same.



6 Summary

In this summary we want to mention all software components and necessary agreements again, since these have to be subject of the Lund discussions.

6.1 Software Components and Certificates

The following software off-the-shelf components will be used within the prototypical system:

- Suse LINUX as underlying operating system for the services
- certificates based on the TERENA TACAR list
- the Handle System to resolve URIDs
- Open LDAP for authentication
- Shibboleth for the exchange of authorization information
- Apache as first component handling http requests

The exact versions have to be defined and upgraded in the official Definition document. Partners who deviate from this have to take care of their local adaptations.

The following software components have to be developed within the DAM-LR project and will be part of the prototypical system:

- a Resource Manager as described above
- a lookup routine that extracts from the LDAP root which partners are federation members[8]
- the MPI intends to extend its ANNEX and LEXUS applications to be used as test beds for the multi-resource scenario
- an Authorization Handler that interacts with other components in the way described above and that provides the necessary forms and process facilities
- a Mover/Copier that takes care of URID database modifications (MPI already started building this component)

The requirements for the components have to be discussed and specified and the work has to be distributed in Lund.

[8] When using a root node for LDAP someone has to house it. MPI will do so, but others can do as well. Still it may be that other architectural solutions may be chosen.

6.2 Agreements

General

- all final agreements and specifications have to become part of the definitions document
- the timing of the various activities will be specified in another document version
- newly developed components should be designed such that suitable APIs are available to support re-usage and will be open source

Federation

- the partners start to synchronize about the foundations of a federation
- the following technical agreements are part of such a foundation

PKI System

- every partner will start activities to become at least a RA under an accepted TERENA TACAR authority

URID

- the Handle System will be used to manage and resolve URIDs
- every partner is a Handle Authority, i.e. requests a prefix from CNRI and install a Handle Service
- MPI will setup mirror services for all partners (others can do as well of course)
- every partner will specify a syntax for its post-fixes and will make them explicit
- every partner will create proper URIDs and maintain its URID database in a consistent way
- access right information will be associated with URIDs and part of the URID database
- all partners will use the same unified record structure for URIDs including the authorization information (the exact format will have to be specified soon)
- MPI will develop a module for URID database manipulation and specify an API (to become part of the definitions document)

Authentication

- LDAP is the prototype system for authentication, partners can chose their own option but all adaptation work has to be done by them
- the partners agreed on a number of exchangeable user attributes
- the partner agree to carry out user management that will have relevance for DAM-LR in a careful and trustful way
- the partners agree on durations of user and usage entries
- the exchangeable user information will become part of a joint domain that allows federation wide searches
- If it will be chosen to go via a joint LDAP root the MPI will volunteer to set it up and maintain it – other partners can do the same

Authorization

- access handling is done by the originating[9] institution
- the access rights information is part of the URID database of the originating institute
- Shibboleth is used to exchange user information
- every partner will set up Shibboleth services
- Apache is used as entry point to handle HTTP requests, redirection tables are set up by the partners such that metadata is open, but that all resource requests are handled by an appropriately designed resource manager
- a prototype RM will be developed, partners are free to develop their own component, but have to adhere to reliability requirements and carry out careful tests
- the MPI will adapt ANNEX and LEXUS to have test beds for the web-application scenario
- the partners will discuss the requirements for access management (processes, rights, ...)
- a prototypical Authorization Database will be designed, that will be based on the requirements
- for the management of access issues a prototypical Authorization Handler will be developed, which will integrate those requirements that can be implemented given the constraints of the

[9] The originating institution is the one where the original copy of a resource was deposited.

DAM-LR project; partners are free to develop their own component, but have to adhere to reliability requirements and carry out careful tests

- prototypical web forms for requests will be created, partners are free to develop their own component, but have to adhere to reliability requirements and carry out careful tests
- additional sites may be added to the list of Identity Providers for testing if they adhere to the trust conditions

7 References

[ELAR] Endangered Languages Archive
http://www.hrelp.org/archive/

[IMDI] ISLE Metadata Initiative
http://www.mpi.nl/IMDI/

[ISO8859-1] ISO-8859. International Standard, Information Processing, 8-bit Single-Byte Coded Graphic Character Sets, Part 1: Latin alphabet No. 1, 1987

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

[OAIPMH] The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0, 2002-06-14
http://www.openarchives.org/OAI/openarchivesprotocol.html

[OAIS] Reference model for an Open Archival Information System, January 2002
http://www.ccsds.org/documents/650x0b1.pdf

[OLAC] Open Language Archive Community
http://www.language-archives.org/

[RFC1738] Uniform Resource Locators (URL), December 1994
http://www.ietf.org/rfc/rfc1738.txt

[RFC2279] UTF-8, a transformation format of ISO 10646, January 1998
http://www.ietf.org/rfc/rfc2279.txt

[SESSIONS] Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen, 2003
http://www.mpi.nl/IMDI/

[UNICODE] Unicode Standard
http://www.unicode.org/

[WINLATIN1] Windows-1252 codepage (WinLatin1)
http://en.wikipedia.org/wiki/Windows-1252

[XML] Extensible Markup Language (XML) 1.0 (Third Edition), W3C, 2004
http://www.w3.org/TR/REC-xml/

[XMLSCHEMA] XML Schema Part 1: Structures Second Edition, W3C, October 2004
http://www.w3.org/TR/xmlschema-1/

[DAM-LR] Distributed Access Management for Language Resources
http://www.mpi.nl/DAM-LR/

[HS] The Handle System
http://www.handle.net

[Shibboleth] Single Sign-On System, Internet2
http://shibboleth.internet2.edu

Appendix A

DAM-LR as a Language Archive Federation: strategies and prospects

David Nathan and Remco van Veenendaal

SOAS, London; INL, Leiden
djn@soas.ac.uk

1. Introduction

The DAM-LR partners are well on the way to forming a federation. The term 'federation' has at least two quite different meanings and it is important not only to distinguish them but also to put our own stamp on what kind of federation we create.

'Federation' has a specialised meaning in information technology, referring to bringing information resources together via information management and networking techniques. It also has an organisational meaning, referring to agencies and their aims and strategies for collaboratively dealing with identities, resources, and services. In this paper, we refer to 'federation' in the first domain as federationT ("federation technologies") and in the second domain as federationA ("federation agency/-ies").

2. FederationT: a background

Federations in the IT sense go back to the earliest days of electronic networking. For example, in 1967 the Online Computer Library Center (OCLC, http://www.oclc.org/) at Ohio State University started sharing bibliographic authority files with fellow libraries, and has long been involved with the issues that still face us now: standards, metadata, quality levels, technology, membership and collaboration. The OCLC now has 9,000 members who share 65 million records to assist in their librarianship work. By the mid 1990s the term 'federated database' was well known. Dempsey et al, for example, describe a "federating solution … [that] allows services to develop autonomously while projecting a single unified image to the user". The motivation for federating resources is to provide value to users by providing a larger metadata set with a correspondingly greater ability to "relieve … potential users of having to have full advance knowledge" of the existence or nature of resources (Dempsey et al). According to this definition, search engines such as Google also, in a weak sense, federate all the web pages that they index.

Lynch (1998) refers to Dublin Core (DC) – also with its roots in libraries – as a tool for federating existing resources: "networked information discovery and retrieval [through] federating disparate, independently maintained databases … [via] a common semantic view of the various databases involved". DC was intended to enhance resource discovery in an open networked world, i.e. in a world without librarian operated-catalogues where quality and consistency are principal values and practices. Dublin Core provides, then, a low-density, lowest-common-denominator but unified method for description and discovery in a unified search domain (the www) by a wide variety of professionals, data-creators and ordinary users. To achieve this, the DC consortium dealt with issues such as (i) syntactic standards e.g. for data and metadata encoding (ii) semantics, e.g. ontologies, semantic web etc. (iii) strategic goals, e.g. selection of the lowest-common-denominator approach to lower the costs and other barriers to coding.

Note that computing power here is a catalyst rather than a central factor; most of the activity is done by humans. FederationT in the sense discussed here contrasts with its use elsewhere to refer to linking networks or grids of computers in order to provide a scaling up of computational power. Here, we seek to scale up resource discovery, retrieval, and preservation, rather than processing.

More recently, parts of the linguistics community have been working in similar areas – OLAC, which was similarly centred on strategic goals for resource discovery, and GOLD ontology, which focussed on mapping out the concept territory of linguistics, to enable linguists to cross-map their varied terminologies (i.e. to bridge between author-created metadata and unified metadata formalised by a body of professionals). OLAC has been moderately successful, although more in terms of raising awareness about issues in language data handling than in unifying resource discovery across language data repositories, possibly because of its broad but ambiguous ambit ranging from endangered languages to multimedia to any language data. GOLD has been motivated by the putative needs of the "endangered languages community" (http://emeld.org/workshop/2003/paper-terry.html), but has mainly drawn interest from typologists and computationalists.

Ultimately, resource discovery has not, at least so far, been a foreground problem for most linguists. In other areas, web search engines have provided alternative solutions, and various areas of industry and commerce have been unobtrusively implementing EDI systems.

A conclusion one might warily draw is that the linguistic community has not (at least yet) found a clear need for such resource discovery and ultimately federalism among repositories. On the other hand, however, linguists will benefit from previous and current work when the day comes that they do find such needs. Progress is likely to be sudden rather than evolutionary, when, at some point, linguists find that not only their tools (email, word processors, databases) but also their modes of expression are electronic (most likely this will occur among the forthcoming generation that will have been fully imbued with electronic communications of all kinds). Once enough linguists' decide to disseminate their own resources via electronic repositories, then federated electronic repositories will become a major locus for searching for other linguistic materials.

3. Opportunities

The current environment for language and technology and the nature of the DAM-LR partners suggest a number of opportunities that can guide strategy for collaboration. Our archives have relatively clear conception of our aims, holdings, and audiences, enabling us to exploit the valuable insights from specific linguistic (and related) subdomains, such as specialised corpuses, endangered languages, sign languages, the collection and implementation of protocol, new genres of data and presentation, new modes of access, and recognition of the new client groups for whom language data is crucially important.

Federating offers us important opportunities, because our repositories hold data that is typically fragmented, not published (or not conventionally publishable), and rare (in fact, it is the fragmented, data-oriented nature of our materials that unifies them as much as the fact that they are linguistic resources). Federation will provide increased dissemination opportunities and therefore add value to our individual collections.

In addition, we have a focal client group, depositors, to whom we need to offer substantial services in order to live up to our manifesto for "Live archives" (DAM-LR). While we do see depositors as a class of archive users, depositors have particular needs, for example to prepare and maintain their materials. The kind of interoperability typically provided by federation is based on use of a single SQL-like query to interrogate multiple repositories, which is centred on the information seeker rather than the information manager, which depositors are becoming. MPI's Lamus is a tool that is offering support in this direction. Another concern of depositors, to attain recognition of archive deposits as significant intellectual contribution on par with conventional publishing, can be greatly aided through successful federated dissemination of materials.

Finally, federation allows us to pool and share our strengths, for example, MPI's IMDI infrastructure and programming strengths, INL and Lund's corpuses, and SOAS' expertise in endangered languages.

4. Federating the domain

The goal of federationT is interoperability, the effectiveness of which is traditionally evaluated by the information retrieval measures precision and recall. Precision and recall are improved by using constrained metalanguages. The more lowest-common-denominator the approach to descriptive metadata (and therefore federationT), the less the specialities of participating agencies are reflected. For agencies that wish to serve users more thoroughly, metadata that drives resource discovery needs to be richer and domain-oriented. However, the mere sharing or overlapping of domains does not guarantee a shared semantics or vocabulary. Colomb (1997) shows that inter-database semantics or metadata mapping is a significant problem, even for simple domains. Agents within a federationT are faced with problems of semantic heterogeneity across their databases. Semantic heterogeneity can be a result of differences not solely between data categories, but between participant's understanding of their meanings, interpretations or usages (Sheth and Larson 1990, quoted in Colomb). It can be about differences in formal data models, system or project goals, or as a result of evolution of these over time.

Language archives face quite different data semantics from business and industry. Business data is anchored in well-defined concepts such as quantification, currency, and product codes; these are clearly-understood abstractions, widely agreed to represent key attributes and whose relation to the real world are not subject to interpretation. Libraries also enjoy conventionality of most of their descriptive attributes: well-understood concepts of author, title etc; in addition, these data are typically provided by authoritative publishers, and, as mentioned above, are available to individual libraries from centralised bibliographic sources.

In this sense, the language data world is a quite distinct one, with its descriptive categories, rather than being predetermined and centrally provided, needing to be derived bottom-up from our widely

varied data and methodologies. A nomenclature of linguistics exists, but language data does not consist of measurements or key attributes, but speculative and contestable interpretations.[10] Thus, the apparent paradox that linguistics seems to guarantee non-interoperability arises due to the nature of language data (which is already metadata, i.e. we do not have agreed-upon data that will "ground out" the metadata semantics), and due to other factors such as that human languages are different from each other in arbitrarily complex ways and that individual linguists seek to emphasise or differentiate aspects of their data or analysis.

Repositories can federate with varying degrees of retention of their "design autonomy" (Colomb), i.e. different levels of change to their information systems to meet the needs of the federation. This is an important issue for DAM-LR. While all the partner agencies hold language data with common but specialised characteristics (e.g. sensitivity; identifying particular persons; emphasis on sound/video in binary formats), they are nevertheless quite specialised. Indeed for most it is a central mission to make a distinct contribution, manifested by creating new infrastructures (e.g. IMDI in the case of DoBeS); others (such as INL) have areal specialisation, or, like ELAR at SOAS, policy specialisation such as collection and implementation of protocol data. In addition, the nature of linguistic data itself is changing and diverging rapidly as the new paradigm of language documentation (a response to language endangerment) grows. For DAM-LR, some concepts are likely to be especially difficult to unify across partners, especially those related to granularity, such as the meanings and cross-mappings of bundle, collection, session etc., and categories of access rights.

5. FederationA: organisational and strategic aspects

The key to dealing with the issues in the preceding section is that the standardisation that enables federationT "is not primarily a computing process" (Colomb); it requires people-based structures, communication channels, and significant resources to maintain these and to enable these to be harnessed towards effective and ongoing development. It is the task of these federationsA to create and host an ongoing, evolving universe of negotiation, knowledge models, and transactions, not merely technical interoperability of terms.

Agencies aiming to form a federation need to be clear about a number of matters, from the semantic ones discussed above, to their purpose and scope, membership, and other strategic, organisational and legal questions. Purpose and scope could range from very broad[11] – to very narrow e.g. 17th century American visual culture (Ninch 2000). These in turn help to create informed and realistic user expectations; i.e. the federationA aims must provide both a forum for sharing and negotiation and a vehicle for disseminating. A co-ordinating body is needed to provide this forum, and to make decisions and strategy, especially in a period of rapidly advancing technology, and where the technology influences what services are expected and provided to users, and have significant financial implications for members.

Therefore, the core of federationA consists of a membership, and its goals. This is totally unlike perceptions of a federationT that consist only of technical standards broadcast from a central agency (the same lesson was learnt in the early development of Z39.50). One could go as far as requiring some form of membership even for users, who must ultimately (for a specialised domain) become part of the community of understanding of the federated metadata and its relationships.[12]

We do have special concerns. For example, conventional authentication systems (such as Shibboleth) exchange minimal data about users, and leave detailed gatekeeping up to individual repositories handling access. However, many linguistic resources have access conditions that associate resources with users, rather as if particular books in a library are not only borrowed under different terms by staff and students, but may be only borrowable by particular named individuals. In our specialist and changing area, federation is not only about searching multiple repositories but about identifying a range of user groups and their needs. This in turn will be enhanced by experience and feedback that a federal forum can incorporate into ongoing strategy.

6. Resourcing and legal aspects

Federation inevitably involves standards, which means formulating rules about implementing them, and, in turn, enforcement through either "incentives or penalties" (Colomb). The mechanism of membership needs to be clear, so that members are signatories to relevant statements of practice, with and formulations of what counts as compliance. Depending in the scope of its activities, a Federation

[10] For example, a transcription might be changed as the linguist better understands a language's structures. Chomsky's aim was to lay foundations of a linguistic theory that would ground out this problem but it has not been overwhelmingly influential in our areas.

[11] Which can raise problems, such as OLAC's adoption of DC-type scope while appealing to language endangerment for its motivation, thus diffusing its clarity of purpose.

[12] Although we should try to avoid the abuse that the term 'community' currently suffers, such as the "Windows user community" or the "open-source community".

might also be responsible for compliance (and reporting) with various legal requirements (such as data protection, privacy etc.) on behalf of members. These various requirements – heightened by the specific sensitivities and potencies of our holdings – mean that initial statements about trust, ethics etc need to be roundly discussed and formulated as a code to which members assent.

Some of our specialisations create limits to the extent that repositories can be federated. For example, one way of making two data sources comparable is to lose some specificity of the more constrained field – i.e. a "lossy" merge that nevertheless allows users to retrieve the relevant data under most queries. However, where a data attribute has legal or ethical implications (e.g. related to intellectual property, access restrictions, or privacy), then the option to manipulate the appearance, content, or granularity of such data is not open. In this example, one can see that ultimately federationA is inseparable from federationT, because technologies must reliably implement the policies of members as legal entities with legal and ethical responsibilities and liabilities.

A federation will need a forum or body that can answer the questions that a legal mind will ask; questions such as: Who owns what? What are the risks and who is responsible for them? Where are the boundaries between agencies? How are differences across jurisdictions handled? Who is accountable? Who can communicate on behalf of the federation? For example, privacy legislation require that someone meet an individual's requests to examine data held about them, which would need to be handled initially at the same level as the "seamless interface" that federationT implies to the wider world. Ultimately, such legal and organisational aspects probably need be formally modelled and integrated into the implementation – again, we see the co-dependence of federationA and federationT.

The activities described in sections 4 and 5 above cannot take place without resources. However, in some cases, the resource base can be hidden or go unnoticed; for example, where participants are (a) performing tasks that are part of their core remit i.e. for which local resources can legitimately be expended; (b) public institutions such as libraries that are expected to develop public infrastructure; or (c) in a homogenous, stable, and well-integrated domain, so that benefits from investment could reasonably be assumed to accrue to all participants. Many of these conditions do not hold for the DAM-LR partners and their domains. Therefore, developments are dependent on obtaining sources of funding together with negotiations about the dedication of local members' resources to the federation's benefit. Again, this will place constraints on the processes for membership.

People resources are also needed: Ninch suggest that a federation may need access to a number of types of skills not only on the IT side (e.g. systems analysis, user interface, programmers) but also linguistic, archive, IP and legal experts, representatives of user groups.

7. Conclusion

DAM-LR is providing a useful testbed for the development of a federation of language resource archives, which could be extended to other nascent groups, such as DELAMAN. It already meets several of the considerations discussed above; in particular, we (i) have clear and constrained tasks and membership; (ii) there is a project and funding scenario within which our tasks are negotiated and resourced. On the other hand, it would be misleading to ignore the diverse and distinct organisational, strategic and implementation issues, and to conflate them all under the one term 'federation'. This paper has shown that a federation will weave together aspects of federationA and federationT.

The function of a federation, then, is to:

- supply services to particular communities (cf. OAI "designated communities")
- to supply those services from allocated resources, i.e. federations must *choose* the communities they will serve (for which there needs to be a forum for negotiation and evolution)
- supply services that take advantage of its members' resources, priorities and values
- to manage its membership and resources in support of the above

References

Colomb, R.M. 1997. "Impact of Semantic Heterogeneity on Federating Databases" *The Computer Journal* Vol 40, No. 5, pp. 235 -244.

DAM-LR (partners). 2006 *Live archives*. Pamphlet.

Dempsey, L., Russell, R., Heery, R. 1997. "Discovering Online Resources. In at the Shallow End: Metadata and Cross-domain Resource Discovery." http://ahds.ac.uk/public/metadata/disc_07.html

Lynch, Clifford 1998. "The Dublin Core Descriptive Metadata Program: Strategic Implications for Libraries and Networked Information Access." In ARL 1998 (196), Association of Research Libraries

NINCH 2000. "Federating Digital Image Repositories and Interpretive Information." *http://www.ninch.org/bb/proposals/visual2.html*

Appendix B

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 form each other, with a Public Key Infrastructure [13] used to sign certificates. Every federation member has to apply to their national Certificate Authority to request the status of a Registration Authority, if the appropriate university is not already a Certification or Registration Authority. Once

recognized as a Certification or Registration Authority, sites can issue or request certificates that will be accepted within the EUGridPMA domain.

3.4. Distributed User Authentication

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

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

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

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

3.5. Access Authorization

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

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

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

4. Additional functions and Specific Implementation Issues

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

4.1. Metadata Utilization.

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

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

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

(3) Unstructured search over the whole domain or selected parts of it. Users can enter words or regular expressions into a free text field (Google-like). Any metadata element including the free text descriptions that contains matching

strings will produce a hit. The recall with this method can be expected to be higher compared with structured search however, the precision will be poor.

4.2. Versioning of Resources.

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

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

4.3. Access Management System

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

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

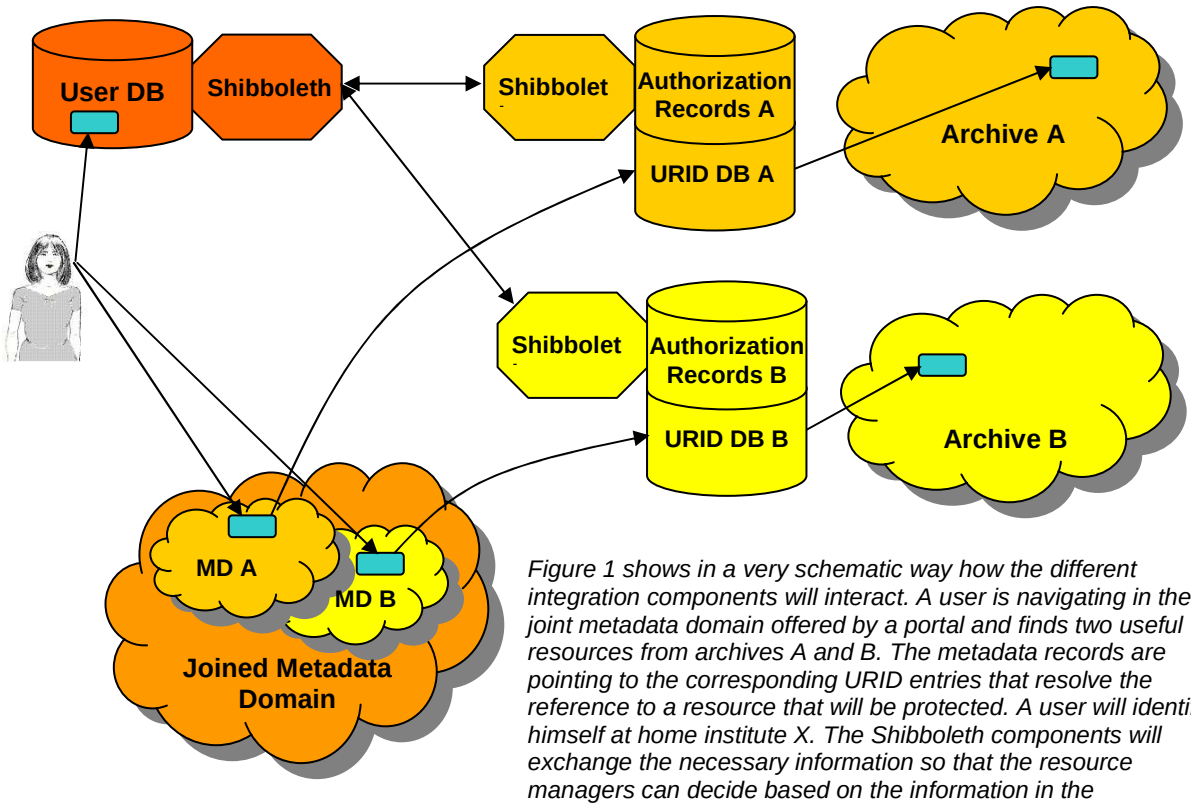


Figure 1 shows in a very schematic way how the different integration components will interact. A user is navigating in the joint metadata domain offered by a portal and finds two useful resources from archives A and B. The metadata records are pointing to the corresponding URID entries that resolve the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

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.

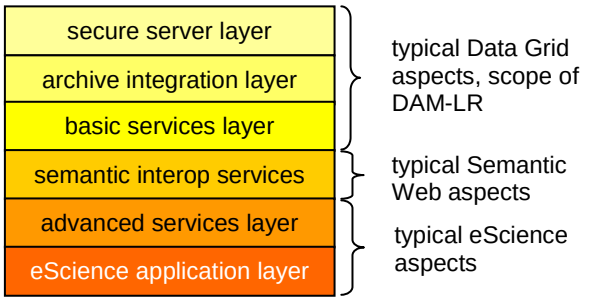


Figure 2 indicates the typical layer hierarchy where Grid solutions take care of typical integration aspects, Semantic Web solutions address the problems associated with interoperability in particular at the semantic level and eScience solutions provide advanced applications such as semantic weaving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

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

6. References

[1] live archives, http://www.mpi.nl/dam-lr/live-archives
[2]GRID forum, http://www.gridforum.org
[3] DAM-LR project, http://www.mpi.nl/DAM-LR/
[3] GTK, http://www.globus. org/
[4] PURL, http://www.purl.org
[5[HS, http://www.handle.net
[6] http://www.mpi.nl/IMDI
[7] Wittenburg, P., Peters, W., Broeder, D. (2002). Metadata Proposals for Corpora and Lexica. In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326
[8] OAI/PMH http://www.openarchives.org/OAI/openarchivesprotocol.html
[9] CNRI. http://www.cnri.net
[10] TACAR. http://www.tacar.org/
[11] EUGRID, http://www.eugridpma.org/
[12] PKI, http://www.pki-page.org
[13] http://shibboleth.internet2.edu/
[14] http://www.science.uva.nl/research/air/projects/aaa
[15] http://www.openldap.org