Deliverable 8.1
Definition 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: D8.1

Authors: INL, MPI

Responsible: INL

Date: 18.01.2007

Content

1 General Agreements

The following specifications were agreed upon at the kick-off meeting and through further interaction between the DAM-LR partners.

1.1 Data formats

For all structured data to be exchanged [XML] will be the exchange language.

1.2 Data encodings

All character encoding is in [UNICODE].

1.3 Software Implementation

For software development Java SDK 1.4 will be used (see [JAVA]).

1.4 PKI System

The project needs a Public Key Infrastructure [RFC3280] and makes use of agreements that are for example available from the [DEISA] project.

The TERENA TACAR supported list offers a suitable way to build a domain of trusted servers and services. The partners in DAM-LR will become at least RA and will get their certificates from the corresponding national root authorities that are supported by TERENA TACAR:

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

The MPI became RA and received its certificate from GWDG that is accepted Certificate authority under the DFN. The current state of affairs (discussed at a DAM-LR meeting at SOAS in August 2006) is that SOAS has attended the required course for new RAs in 2006 and is well on the way of becoming RA, INL preferred to make use of their mother university's (University of Leiden) RA status and has requested and received certificates and Lund has also started talks with their mother university about becoming RA.

1.5 References

[DEISA] Distributed European Infrastructure for Supercomputing Applications
http://www.deisa.org/

[JAVA] Java 2 Platform, Standard Edition (J2SE), Sun, November 2003
http://java.sun.com/j2se/

[RFC3280] Internet X.509 Public Key Infrastructure, April 2002
http://www.ietf.org/rfc/rfc3280.txt

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

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

[TERENA TACAR] TERENA Academic CA Repository
http://www.tacar.org/

2 Metadata

Every partner is free to choose his own metadata descriptions, however, at the integrated DAM-LR level the [IMDI] infrastructure will be used as specified by the IMDI Schema version 3.0 (see Appendix D). If partners choose another set than IMD,I they have to provide a semantic mapping and a port such that metadata can be harvested. In general for metadata exchange two options are possible: (1) The [OAI] Protocol for Metadata Harvesting [OAIPMH] is supported; (2) A URL (see [RFC1738]) is specified and points to the local IMDI root node. The vocabulary of the set at the distributed level is defined by IMDI semantics. Extensions of the controlled vocabularies were seen as possible. In particular, the type of resources has to be broad enough to cover the whole diversity.
Option 2 was selected as the preferred way for creating the DAM-LR integrated metadata domain.

2.1 References

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

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

[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

3 Unique Resource Identifiers (URIDs)

It was agreed that URIDs have to be introduced and that resource attributes that need to be disseminated to others ,such as rights should be associated with these URIDs, since they are the unique reference for a certain object.

3.1 URID resolving

For resolving URIDs to physical paths the Handle System was chosen (see [HANDLE]). The architecture choices for DAM-LR are as follows:

- 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[1] 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.
- 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 APIs 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 any local server problems the URIDs can still be resolved.

A detailed time schema was discussed in Lund in January 2006. It was updated at SOAS in August 2006.

At a technical meeting at MPI in November 2006, the idea of storing user rights in the Handles was discussed. The main reason for storing access rights in the Handles is being able to copy resources between archives: user rights information must be propagated and safeguarded when copying resources. Although not a specific requirement for the DAM-LR project, the consortium will future-proof the infrastructure by doing some research into this feature.

A first idea for storing user access rights in Handles is to store the rights next to the URLs of the resources. E.g. in the form:

"userID1[2] userID2 … userIDN"

A (copied) Handle could be visualised like this:

Handle	*URL*
	URL of copy
	"userID1 userID2 … userIDN"

Writing will not be allowed on copies resources.

[1] The Handle System created by CNRI is a widely used system so that we can expect reliable services in the future.

[2] For more details about userIDs, see the section on Access Management.

3.2 Naming system

It is recommended not to use semantics within the postfixes, but in fact every partner is free in his decisions. At a meeting at SOAS in August 2006, the partners decided to use the following postfix systems:

- MPI will use the ‘Specification of MPI Resource Identifiers’ as presented in Appendix C
- INL will use the same postfix system as MPI
- Lund: Lund will use the same postfix system as MPI
- SOAS: the SOAS postfix system does include some semantics and is described in detail in Appendix G

3.3 References

[HANDLE] The Handle System, CNRI
http://www.handle.net/

4 Authentication and Authorisation

4.1 AA Systems

It was agreed that a number of solutions for the authentication and authorisation in distributed scenarios will be analysed. In general a modular system was preferred instead of taking a comprehensive and therefore complex solution. It was agreed that sensitive data should not be transmitted across the open networks to not hurt privacy. It was also agreed that authentication should preferably done by the home institutes of the users with whatever system they use (if this is possible). The following solutions were investigated:

- Shibboleth for distributed authorisation [SHIBBOLETH]
- AAA Toolkit for distributed authorisation (and accounting) [AAATOOLKIT]
- Usage of the Handle System (see below) to deal with authorisation issues
- A-Select for multi-layered authentication although this will not be put in use at an early stage [ASELECT]

A careful architectural investigation had been done by the MPI team. It was presented already at the 3rd DELAMAN meeting in Austin in Texas (see PPT file as reference) and a Special Report has been written and has been subject of the discussions at Lund in January 2006 (see Appendix E). From this analysis the following choices are obvious:

- Shibboleth will play the role to exchange open user attributes. The reasons for this choice are that (1) Shibboleth offers robust and secure software that is increasingly often used in similar scenarios and (2) Shibboleth is accepted by increasingly more countries as basis for distributed authorization solutions.
- Since Shibboleth offers standard interfaces to LDAP, since LDAP offers solutions for creating joint search domains which can become important for DAM-LR purposes and since LDAP is widely used by academic institutions for authentication and user management, there are strong arguments to use LDAP at first instance as the prototypical solution. However Shibboleth supports also other user database systems and the choice is probably best guided by existing installations.

Other solutions have still big disadvantages such as the need to develop additional code or to offer not widely tested code. Therefore, they will not be selected for the DAM-LR project.

At a technical meeting at MPI in November 2006, the “[BROWSER/ARTIFACT] SSO profile” of [SAML 1.1], supported by Shibboleth, was chosen as the default profile for DAM-LR.

4.2 Access Management

It was agreed that for DAM-LR we should define the policies and rights that have to be handled in an access management system. Also the type of users and user groups and possible attributes have to be clarified. All partners will make an inventory of access restrictions currently applicable to their language resources. This is the basis for the definition of access policies and rights.

INL has made their access management document available to the other partners via the DAM-LR wiki website. It has been included in this document as Appendix H. SOAS has provided their access management information in the Archive Formation document (Deliverable 4.1 T16). The relevant section of that document is also available as Appendix I of this document. Lund have made their access management information available in the same Deliverable 4.1 T16. The relevant section of that document is available as Appendix J of this document. The MPI access management policy is described in appendix K.

The minimal user attribute set the partners will store and exchange authentication (enabling the single sign-on) will look like this (and is described in more detail in chapter 4 of Appendix E of this document):

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	(code for) hosting institute, that administrates the primary account for the user and where he can be authenticated (Shibboleth)
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 user identification 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)

At a technical meeting at MPI in November 2006 a joint specification for the userID was discussed. The userID will be of the form:

Federation ID + special character + User Identifier

The Federation ID or unique archive identifier will consist of 6 characters and could look like “INL_LA” (INL Language Archive). The special character is – for now – a colon (‘:’). These IDs must be usable as uids in .htaccess files. An example of a userID could be:

INL_LA:myUserId

This example userID indicates that the user with “myUserId” is a user who is known to and/or trusted by the INL.

4.3 References

[ASELECT] A-Select, SURFNet
http://a-select.surfnet.nl

[AAATOOLKIT] AAA-Toolkit, UvA
http://www.science.uva.nl/research/air/projects/aaa/index_en.html

[SHIBBOLETH] Shibboleth, Internet2
http://shibboleth.internet2.edu

[BROWSER/ARTIFACT] SSO profile
https://spaces.internet2.edu/display/SHIB/BrowserArtifact

[SAML 1.1] Security Assertion Markup Language
http://www.oasis-open.org/specs/index.php#samlv1.1

5 Federation

When the DAM-LR partners (virtually) join their local archives a new, virtual organization is created. In proper grid computing [GRID COMPUTING] terminology, they form a federation [FEDERATION]. The DAM-LR partners form a federation and the users of the portal(s) make use of the services the federation provides. New partners may want to join the federation in the future, linking their local archives to the distributed solution.

A first draft of a user agreement for the DAM-LR federation(‘s services) was discussed at a DAM-LR meeting at SOAS, U.K on August 25, 2006. Later versions of this deliverable will contain updates of the user agreement (and e.g. agreements between the DAM-LR partners). The draft user agreement has been included as Appendix F of this document.

5.1 References

[FEDERATION] Federation
Nathan, D. and Van Veenendaal, R. (2006), “DAM-LR as a Language Archive Federation: strategies and prospects”, in: *Proceedings of the LREC 2006 Pre-Conference Workshop*: 27-30, 2006 May 22, Genoa, Italy. Paris, European Language Resources Association.

[GRID COMPUTING] Grid computing
http://en.wikipedia.org/wiki/Grid_computing

Appendix A

DAM-LR Meeting Report, 12/13 July 2005

DAM-LR

Meeting Report, 12/13 July 2005

Peter Wittenburg, Daan Broeder, Freddy Offenga
2005-7-14

Introduction

On the 12th and 13th of July 2005 the first strategic meeting of DAM-LR, including the Executive and the Working Committees, took place at the MPI in Nijmegen. With the exception of Peter Austin, all WC/EC members, the project coordinator and some additional technical staff members were present to join the presentations, demos and initial discussions. This report gives a rough summary of the presentations and discussions. For more detailed information we refer to the presentations.

Project Overview

The meeting started off with a presentation of the main overview of the DAM-LR project. The current situation of scattered resources was hightlighted to show what the problems are, and how they could be solved by integrating the archives using the four pillars of DAM-LR. The tasks were all discussed briefly to get a clear understanding of what has to be done and who is responsible. It became clear that of the involved partners each have different requirements for their archives, and that they all need to think about a working solution at the local level when discussing the aspects of the distributed solution.

All work packages were discussed in detail. As part of the definition task (WP8) INL will describe what decisions are made, e.g. why a certain software component is chosen, which interface specifications were agreed upon etc. The importance of testing (WP11) came up since this must prove that the real system is actually working in a robust way.

Important events like LREC, DELAMAN and E-MELD were mentioned where we have to present papers and gather comments that are relevant for the project. Lund offered to organize training courses by using their excellent training facilities. All partners should ensure that the DAM-LR goals will be presented at various events and send notifications to the coordinator.

All remaining questions about the DAM-LR goals and their implementation could be clarified.

The relation of DAM-LR with the DELAMAN network, and the need to achieve good collaboration was explained and agreed.

Project Management

In the second talk the project management issues were presented to the partners. The members of the EC and WC were presented, and it was pointed out that there will be a strategic meeting of the EC each year and virtual meetings of the WC every three months. Again the relevance of careful reporting to the EC was stressed, and everyone was asked to submit the required documents in time. The MPI will create forms that can be used to achieve a high degree of unification.

All are invited to take a look at the DAM-LR web-site and help to improve it. Everyone agreed that almost all documents created within DAM-LR should be published on the Web-Site. DAM-LR is at the cutting edge of Data Grid technology and should offer the opportunity for others to look at the positive and negative experiences. INL offered to set up a Wiki page to share knowledge and have discussions about all DAM-LR aspects.

Local Prototype

After the break the MPI presented the state of the local prototype solution to access and manage language resources at the MPI. The prototype solution can be seen as a kind of reference solution and it should contain solutions for all four pillars mentioned in the Technical Annex (metadata, unique identifiers, authentication, authorization).

The structure of the archive was demonstrated using a typical IMDI organization where the MPI domain is separated into sub-domains containing multiple levels of corpus structures depending on user needs and requirements. Access to resources is handled at the IMDI metadata level where rights and policies can be defined for users and groups.
The 'resource basket' idea was explained by comparing it to an on-line shopping cart where a researcher selects resources from different archives to create their own virtual and temporary working set. It was mentioned that unique identifiers (URIDs) are relevant right now because other Max-Planck institutes want to offer multiple distributed copies of language resources, and because of the emergence of new types of commentary tools that create all sorts of references to archival resources. URL's can change and they have to be maintained only at one location to make it a tractable job to manipulate them.
The requirements for the local MPI system were presented. A problem with Shoebox lexicon files was mentioned as an example of how specially formated files can depend on other files while users aren't aware of this. There was a discussion about file format checking with custom parsers which would be done at the moment of file ingestion and can only be implemented correctly when the exact structure is known (e.g. a schema).
The usage of archives was discussed. Too many rules could discourage users to ingest data, but a managed archive also helps people because they get free features like automatic backups and data checks. It was agreed that versioning is an important issue. This aspect needs more discussion and still some work has to be done. Format differences were discussed. Sometimes other formats (e.g. mp4 video) will need to be available, these could be stored in the archive next to the main format at the time of ingestion, or converted real-time at the moment when it is needed. The point was made that it is important to have quality control of the format conversion process.
The MPI gave demonstrations of IMDI tree browsing and access management using a web browser. Metadata search was shown using a client tool. The tree copy tool was demonstrated by making a copy from a small part of the archive to the local machine. Next there was a demo and presentation of the Language Archive Management and Upload System (LAMUS) which is a content management system specialized for language resources. Resources and metadata are placed in a user's work space to gather resources, manipulate and prepare them. The resources can be ingested into the archive when the user is satisfied with the created setup and when all checks were without error reports.
The metadata issue in LAMUS was discussed because when using LAMUS only a limited set of elements can be entered and manipulated. People can use the IMDI editor to enter all metadata descriptions and then upload the IMDI file like any other resource. The need for an integrated metadata manipulation interface was discussed, however, it was shown to be dangerous to support a minimal metadata set. Finally the current LAMUS limitations were mentioned and a feature to-do list was shown.
Summarizing, it can be said that the local prototype developed at the MPI fulfils almost all requirements. The only major functionality missing is the support of unique resource identifiers (URIDs). The MPI explained that this topic can now be tackled since all seemed to agree that the Handle System is the only viable alternative at this very moment. In November a complete solution is expected.

Local Lund

After the demonstrations and a short break it was time for presentations of the state of the local archives from Lund, SOAS and INL. Lund started by presenting the organization of common resources from the library, technical group and the many laboratories at Lund University. A great variety of language resources were listed from corpora of Swedish dialects, keystroke logged writings, medieval manuscripts, etc. Several resources are already described using IMDI metadata and some are in the planning for IMDI descriptions. It was mentioned that there's a wish to integrate resources from the 'frog story' domains which could become possible with DAM-LR.
A completely new archive server with 34TB data storage capacity is expected to complete the archive setup in the first months of the coming year. Currently, the data is stored at various servers on the campus. It was mentioned that IMDI metadata at Lund might need to be upgraded and that the MPI can help there. Access to the resources was discussed and the decision was made to document the details.
Lund university imagines it will take over the local prototype solution when it is ready and tested.

Local SOAS

The state of the local SOAS archive was presented. The Endangered Languages Archive (ELAR) is part of the Hans Rausing Endangered Language Project (HRELP) at SOAS where the focus is on digital archiving and dissemination of language documentation. The archive is in the process of being set up and a powerful server system is being installed which will be able to store all data from the various documentation teams is in the process of being installed. There will be 20 projects each year producing about 0.7 TB content data.
Digital data and other data (to be digitized) will be ingested into the archive. The archive will be structured by relying on relational database technology in combination with a file system. An internally used metadata schema will have to be derived from which it will be possible to generate OLAC and IMDI type of metadata records. The records will be offered by using the OAI PMH protocol, and IMDI records will have to be presented using some hierarchical structure to allow browsing. The archive catalogue (metadata) should be open and available through the web.
Other architecture aspects have to be worked out in the coming months. SOAS will document the architectural decisions and inform the others about federated archives.

Local INL

Finally, INL presented their archive status. There are three online corpora (5, 27 and 38 mln words) which have very limited access due to copyright issues. Therefore these can be interesting access management test cases for DAM-LR. These corpora are available only through telnet, no web interface is available, which is enough for the current users. Another corpus is PAROLE which contains 20 mln words and is TEI encoded. Access is done by means of a usage agreement. All queries on the corpus are logged for security reasons.
A special new centre, the TST-centre, was setup to manage and distribute digital language material which contains the Dutch Spoken Corpus (CGN) and several other important corpora from INL. An IMDI portal is foreseen for this year. The TST is in the process of defining its architecture. All decisions will be documented for the DAM-LR purposes.

Distributed solution - URIDs

The second day started with a presentation of all issues concerning the distributed solution which is the core work of DAM-LR. The goal is to have federated archives which can be achieved by implementing the four pillars of the project (joint metadata domain, joint URID domain, joint user domain, distributed authorization mechanism). From the user perspective it would mean that federated archives are easy to access by having a single sign-on system (SSO) and a transparent view on the language resources.
The first important pillar, URIDs, was presented and discussed in detail. The reasons for using unique identifiers for each resource were explained. URIDs are not tied to physical storage locations, but to name an archival object. Therefore they can be used to identify distributed copies of resource objects. Since these identifiers always need to be resolved it is crucial that resolver systems are robust and always operational. A secure resolver also requires a PKI system to manage the URIDs.
A good candidate resolver system is the Handle System (HS) from CNRI. This system was installed and tested by the MPI and presented in more detail. The HS is already used by several important parties like the Library of Congress, the Defense Technical Information Centre (DTIC) and the International DOI Foundation (IDF). The system uses a handle prefix to indicate the naming authority of the identifier. This prefix must be registered at CNRI. The remaining part of the identifier (Unique Local Name) must be created by the naming authority.
There was a discussion about the separation of the identifier within DAM-LR. The first option is to have one prefix for the project as a whole (or even beyond that to cover the DELAMAN archives). The second option is to let each partner have their own prefix. A point was made that it could be problematic when an institute wants to withdraw from a ‘prefix’. It was noted that the HS can be used to handle versioning where one handle can be resolved to different versions of the same object. Since it has to be possible to refer not only to the most recent version, but also to specific (older) versions, it was decided that every version needs to get a separate URID. A similar discussion came up about web resources whether or not each little object (e.g. a button picture or a linked html) needs an URID. A scenario was pointed out in which there is limited access to a picture on a html page so that the picture would need a URID to set access rights.

The partners agreed to discuss the requirements again via the Wiki site and reach a positive final decision about the scope of the URID domain and the construction of URIDs. No alternative was mentioned for the Handle System, i.e. the MPI will continue with its tests and integrate the Handle System in the local prototype during the next months.

Distributed Solution - AAI

Other important pillars of the Distributed Solution are about the Authentication and Authorization Infrastructure (AAI). Partly they collapse in one system, but in modern IT systems they are separated. These two topics were presented and discussed in great detail. The requirements an AAI were listed and discussed first. It was agreed that the rights and policies set for a certain resource must always remain as set by the owning and originating archive. It is up to the individual archive how data depositors can influence the policies and access rights. Allowing each others users access is an important aspect since federated archives want a single entry point for each user and reduce user management as much as possible. Users should not have to authenticate for every resource since that would make the system unusable.
Shibboleth, a possible candidate component for AAI was presented. This system includes a privacy-preserving negotiation about attributes used for authorization of resources. The main idea is that users are in control of their own private and sensitive information at all times. It is not clear whether or not communication between the service provider (where the resources are managed) and the identity provider (where the users are managed) is always required for every access operation on the resource. In other words, is there a notion of a session and if so, where is the session information stored?
A list of open questions revealed that many aspects of the system are still unclear. Because of the importance of the AAI for DAM-LR all agreed that we need correct and very detailed information about the features of Shibboleth. It was agreed that the requirements and question lists will be discussed and completed via the Wiki, and that the MPI will start communicating with the developers of Shibboleth and users of the system. The other partners will also check whether Shibboleth is used already in their neighborhood.
A-Select from SURFNet was mentioned as a professional authentication system which could be used together with Shibboleth for example. However, it was agreed that the introduction of an elaborated system such as A-Select would not be a very high priority.
During the discussion the question was raised as to what the granularity of the authorization information should be. If every individual has to be handled separately then attributes are not a very powerful mechanism. It was agreed that it is not possible to exchange the password files, for example, amongst the archives, since this is forbidden by national laws and very insecure anyhow. An alternative would be to exchange, for instance, the users email addresses as a simple alternative for authorization.
The partners agreed to look around at their institutes to see which AAI systems are currently operational, get in touch with the responsible persons to get more information and pass it through to the rest of the DAM-LR members. All will try to document their findings about AAI and Shibboleth before the DELAMAN workshop in November where all these issues will be discussed in more detail.

Management Decisions

The partners agreed that we can put (almost) all final DAM-LR documents and relevant reference material on the website. The partners will organize a workshop for LREC 2006. In October there will be a technical meeting (probably in Lund) about the state of the investigations with respect to the Handle System and the AA Infrastructure. With respect to the Handle System final decisions are expected. With respect to the AAI further details will be discussed.

Action Points

- prepare a first version of the definitions deliverable (INL and MPI)
- collect documents relevant for DAM-LR, send to coordinator (all)
- collect relevant links, send to coordinator (all)
- put powerpoint presentations online (MPI)
- put relevant documents and links online (MPI)

- set up wiki pages (INL)
- prepare ideas for training courses and open a Wiki site (Lund)
- gather arguments for the specification of URIDs (all)
- document local systems (Lund, SOAS, INL)
- document local AAI solutions (all)
- inform the partners about federated archives (SOAS)
- establish communication with Shibboleth responsibles (MPI)
- discuss DAM-LR issues for DELAMAN, in particular with respect to the AAI, before November (all)
- prepare for LREC workshop (MPI)
- check the possibilities for a PKI system for all servers (MPI)

The presentations given at the strategic meeting will be put on the web site.

Glossary

AAI	Authentication and Authorization Infrastructure
CGN	Dutch Spoken Corpus
DAM-LR	Distributed Access Management for Language Resources
EC	Executive Commitee
HS	Handle System
HSM	Hierarchical Storage Management System
IPR	Intellectual Property Rights
LAMUS	Language Archive Management and Upload System
PKI	Public Key Infrastructure
TEI	Text Encoding Initiative
URID	Unique Resource Indentifier
WC	Working Commitee

References

A-Select	http://a-select.surfnet.nl
DAM-LR	http://www.mpi.nl/DAM-LR/
DELAMAN	http://www.delaman.org
DOBES	http://www.mpi.nl/DOBES/
ELAR	http://www.hrelp.org
E-MELD	http://emeld.org
HS	http://www.handle.net
IMDI	http://www.mpi.nl/IMDI/
INL	http://www.inl.nl/
LREC	http://www.lrec-conf.org
Lund	http://www.ling.lu.se/
MPI	http://www.mpi.nl
OLAC	http://www.language-archives.org
PAROLE	http://www.inl.nl/eng/corp/parole.htm
Shibboleth	http://shibboleth.internet2.edu
SOAS	http://www.soas.ac.uk/
SRB	http://www.npaci.edu/DICE/SRB/
SURFNet	http://www.surfnet.nl/info/en/
XML	http://www.w3.org/XML/

Appendix B

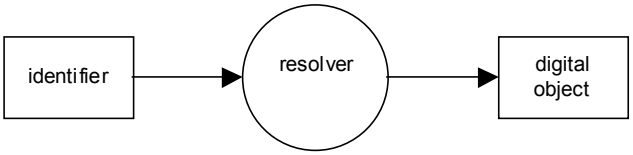
Persistent Resource Identifiers

Persistent Resource Identifiers

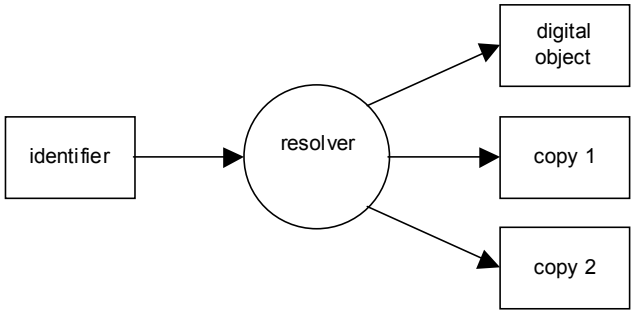
Freddy Offenga, July 2005

1 Introduction

When locating a specific digital object there must be a way to associate a unique name (identifier) with the physical location of the object. The process of associating a name with a physical location is called 'resolving' and a system able to execute such a process is a 'resolving system'. For example, the identifier *http://www.mpi.nl/something.html* is associated with a digital object (a web page) which is stored somewhere on a server at the MPI. In this case the resolving system includes a web-browser, internet protocols and web-servers.



To allow efficient access to digital objects (e.g. movies) there can be several copies of the same object stored at different locations. In such a distributed system a unique identifier can therefore be resolved to more than one location.



2 Requirements

To understand what's a good persistent and globally unique identifier it's helpful to take a look at the specification of [RFC1737] 'Functional Requirements for Uniform Resource Names':

> Global scope: A URN is a name with global scope which does not imply a location. It has the same meaning everywhere
>
> Global uniqueness: The same URN will never be assigned to two different resources
>
> Persistence: It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name
>
> Scalability: URNs can be assigned to any resource that might conceivably be available on the network, for hundreds of years
>
> Legacy support: The scheme must permit the support of existing legacy naming systems, insofar as they satisfy the other requirements described here. For example, ISBN numbers, ISO public identifiers, and UPC product codes seem to satisfy the functional requirements, and allow an embedding that satisfies the syntactic requirements described here

Extensibility: Any scheme for URNs must permit future extensions to the scheme

Independence: It is solely the responsibility of a name issuing authority to determine the conditions under which it will issue a name

Resolution: A URN will not impede resolution (translation into a URL, q.v.). To be more specific, for URNs that have corresponding URLs, there must be some feasible mechanism to translate a URN to a URL

3 Naming Schemes

Already a great amount of schemes exist to describe globally unique identifiers (see [IDDC]). Many schemes follow the Uniform Resource Identifier syntax [URI] for identifying digital or physical resources. The features of the major schemes are summed up in appendix A and the global syntax is described in Appendix B. A short description of each follows in the sections below.

3.1 URN

The Uniform Resource Name [URN] scheme describes identifiers using the URI prefix 'urn:' to define namespaces which are managed by [IANA]. For example, books can be identified using the ISBN namespace like this:

urn:ISBN:<ISBN-number>

Instead of namespace the term 'Naming Authority' is often used to indicate that an authority is responsible for the creation and management of their own (locally) unique identifiers. To make sure that the naming authority itself is unique there has to be one central authority managing the naming authorities (e.g. IANA).

3.2 PURL

A Persistent Uniform Resource Locator [PURL] is simply a URL with a fixed location prefix linking to a resolving service at the OCLC. With the PURL software one can manage and create PURLs.

3.3 Handles

The Handle System is a general-purpose global name service that allows secured name resolution and administration over networks such as the internet. The Handle System manages handles, which are unique names for digital objects and other internet resources [RFC3650].

A handle is specified by:

HNA/HLN

where:
HLA is the Handle Naming Authority
HLN is the Handle Local Name

The naming authority is a number assigned by the handle system administrator (CNRI) in the global handle registry. Each naming authority can further specify sub-authorities as a tree structure, e.g. 10.1045 is the D-Lib magazine under the DOI project.

3.4 ARK

The approach of the Archival Resource Key (ARK) naming scheme is based on the observation that persistent identifiers should be *actionable*, meaning that they should be linked to services that provide persistence [ARKCDL]. Unlike URNs and Handles the ARK identifier defines a special kind of URL which links to the object, the object metadata and a commitment statement from the current provider. An ARK identifier is specified by:

http://NMAH/ark:/NAAN/Name

where:
NMAH is the Name Mapping Authority Hostport which is the current service provider for the ARK identifier. This part is replacable.
NAAN is the Name Assigning Authority Number, a globally unique number indicating the authority responsible for the assignment of names (NAA).
Name is the name assigned by the Name Assigning Authority

4 Specification of Local Names

Persistent identifiers are meant to exist forever and once defined they can not be changed. When any part of the identifier needs to be changed, e.g. the author or subject information, the identifier as a whole is no longer valid. Therefore it's good practice to leave as much semantics as possible out of the identifier string (see [COOL]). A naming authority can use an identifier generator to create unique opaque identifiers. There are tools available to create short unique strings [NOID] which must be managed by the naming authority. In general there has to be some central service which provides identifiers for use within the naming authority namespace.
A second option is to use a tool to generate 'Universally Unique Identifiers' [UUID] (also called 'GUID'). A UUID is a 128-bit number created by algorithms which minimise the chance that an identical UUID will ever be created. Therefore a UUID doesn't require a central authority for administration.

5 References

[ARKCDL] John A. Kunze, Towards electronic persistence using ARK identifiers, California Digital Library, July 2003. http://ark.cdlib.org/arkcdl.pdf

[IANA] Internet Assigned Numbers Authority, http://www.iana.org/

[IDDC] Guidelines for using resource identifiers in Dublin Core metadata and IEEE LOM, http://www.ukoln.ac.uk/metadata/dcmi-ieee/identifiers/

[COOL] Cool URIs don't change, http://www.w3.org/Provider/Style/URI.html

[NOID] Nice Opaque Identifier, http://www.cdlib.org/inside/diglib/ark/noid.pdf

[PURL] Persistent Uniform Resource Locator, http://purl.org/

[RFC1737] Functional Requirements for Uniform Resource Names, December 1994, K. Sollins, L. Masinter, http://www.ietf.org/rfc/rfc1737.txt

[RFC3650] Sam Sun, Larry Lannom, Brian Boesch, Handle System Overview, Internet Engineering Task Force (IETF) Request for Comments (RFC), RFC 3650, November 2003. http://hdl.handle.net/4263537/4069

[URI] Uniform Resource Identifiers, http://www.ietf.org/rfc/rfc2396.txt

[URN] Uniform Resource Names, http://www.ietf.org/rfc/rfc2141.txt

[UUID] Universally Unique Identifiers, http://www.ietf.org/rfc/rfc4122.txt

Appendix A : Features

	URL	**PURL**	**URN**	**Handle**	**ARK**
location independent	No	No	Yes	Yes	Yes
protocol independent	No	No	Yes	Yes	Yes
resolve to multiple copies	No	No	Yes	Yes	Yes
resolution system	http	http	DDDS (draft)	Handle System	http
management tools available	web server tools	PURL Software	-	Handle System	-
naming assignment authority	ICANN	OCLC	IANA	CNRI	CDL

Appendix B : Syntax

Scheme	**Syntax**
URL	<protocol> <host> <local name> example: *http://foorbar.org/resource.txt*
PURL	“http://purl.org/” <name> example: *http://purl.oclc.org/OCLC/PURL/FAQ*
URN	“urn:” <namespace identifier> “:” <local name> example: urn:isbn:1400052939
Handle	[“http://hdl.handle.net/”] <handle naming authority> “/” <handle local name> examples: *10.1025/foobar* *http://hdl.handle.net/10.1025/foobar*
ARK	[“http://” <NMAH> “/”] ark:/” <NAAN> “/” <name> [“/” <qualifier>] examples: *ark:/12025/654xz321* *ark:/12025/654xz321/s3/f8.05v.tiff* *http://foobar.org/ark:/12025/654xz321/s3/f8.05v.tiff*

Appendix C

Specification of MPI Resource Identifiers

Specification of MPI Resource Identifiers

Freddy Offenga, October 2005

All resources in the archive get a unique resource identifier specified by the Handle System [RFC3650] as:

<HNA> “/” <HLN>

where:
<HNA> is the Handle Naming Authority
“/” is a marker to separate the HNA and HLN parts
<HLN> is the Handle Local Name

Handle Naming Authority

A number created and administrated by CNRI. Not yet available.

Handle Local Name

The MPI HLN is a string of alphanumeric characters and separater marks. The syntax is specified as follows:

<Extra> “-“ <Local ID> “-“ <Check Digit>

<Type> := <Digit><Digit>
<Local ID> := <DBlock> “-“ <DBlock> “-“ <DBlock> “-“ <DBlock>
<Dblock> := <Digit><Digit><Digit><Digit>
<Check Digit> := <Digit>
<Digit> := { ‘0’-‘9’ | ‘A’-‘F’ }

Extra

Eight bits (two hexadecimal digits) reserved for extra information.

Local ID

The local ID is a 64-bit number for a capacity of 2^{64} objects. For readablility the number is divided into four groups of four hexadecimal numbers.

Check Digit

For error detection a single check digit is provided which is calculated over the preceding digits (Resource Type and Local ID). The algorithm in ISO Check Character Systems Mod 17-16 [ISO7064] must be used to calculate a correct check digit for new identifiers.

Example

An example of a valid MPI HLN is:

00-0123-4567-89AB-CDEF-x

where x is the check digit calculated according to ISO7064.

References

[ISO7064] Data Processing – Check Character Systems, ISO, Mod 17-16, 2003

[RFC3650] Sam Sun, Larry Lannom, Brian Boesch, Handle System Overview, Internet Engineering Task Force (IETF) Request for Comments (RFC), RFC 3650, November 2003.
http://hdl.handle.net/4263537/4069

Appendix D

IMDI Schema Version 3.0.4

<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v4.2 U (http://www.xmlspy.com) by Daan Broeder (Max-Planck Institute for Psycholinguistics) -->
<!--

VERSION 3.0.4

DATE 4-1-2005
imdi:LanguageIdType also support for ProfileAttributes.
imdi:IntegerType also support for ProfileAttributes.
Session.Date also support for ProfileAttributes
reintroduced VocabularyDef for CVs

DATE 21-12-2004
Changed type of attributes ResourceId from xsd:ID to xsd:string

DATE 09-11-2004
Added minOccurs="0" to Catalogue.Format.Text, Catalogue.Format.Audio, Catalogue.Format.Video
Added minOccurs="0" to Catalogue.Quality.Audio, Catalogue.Quality.Video

DATE 28-10-2004
Changed type of attributes ResourceRef, ResourceRefs from xsd:IDREFS to xsd:string

DATE 20-09-2004
Changed default value for Type attribute of Vocabulary
from OpenVocabularyList to OpenVocabulary

DATE 01-09-2004
Added Catalogue.DocumentLanguages
Added Catalogue.SubjectLanguages
Added SimpleLanguageType
Added SubjectLanguageType

DATE 25-08-2004
Adapted imdi:DateType to have a possible range | empty string

DATE 21-08-2004
chenged type of Catlogue.Authors to support Profile attributes

DATE 09-08-2004
Added CATALOGUE.Profile to possible MetatranscriptTypes
Changed errors in type declaration of Catalogue.Format.Text ... Catalogue.Format.Video, Catalogue.SmalledtAnnotationUnit
Changed typedefs for many elements of Catalogue.Publisher, Catalogue.Size, Catalogue.Pricing

DATE 03-08-2004
Changed DateType of Session.Date to DateRangeValueType

DATE 19-07-2004
Added Keys to CatalogueType

DATE 15-06-2004
Double extension of ProfileAttributes removed from elements with a Vocabulary and from elements with a DateType
Added minOccurs="0" to Project.Description
Added minOccurs="0" to References.Description

DATE 14-06-2004
LanguageId can be Unspecified or Unknown
Source can have both CounterPosition and TimePosition to accomodate Profile(s)

Source can have Profile attributes
LexiconResourceBundle: NoHeadEntries, NoSubEntries to IntegerType (also Unknown & Unspecified)

DATE 13-06-2004
introduced IntegerType for [0-9]+ | Unknown | Unspecified
this can replace CounterPositionType
and also serves for ValidationLevel

DATE 03-06-2004
Some changes: to legalise existing profile practice and regularize things
(BUG)AgeValueType ';' to '.' (See CHILDES)
made LanguageId element in WrittenResource of type Vocabulary
Age and BirthDate can now have ProfileAttributes
Anonyms.ResourceLink can have profileAttributes
Removed VocabularyDef and VocabularyDefType - these were not used
Added FollowUpDepend to profile attribute group

VERSION 3.0.3
DATE 17-05-2004
Added minOccurs="0" to History element
Added required attribute "Name" to NamedLinkType (for CorpusLink)

VERSION 3.0.2
DATE 05-11-2003
History element introduced
Profile attribute of Metatranscript element introduced

VERSION 3.0.1
DATE 20-10-2003
Actor.Age can now also be a range
Added BirthDate to Actor

VERSION 3.0.1
DATE 15-10-2003
Corrected TimepositionType, last occurence of ':' should be optional

VERSION 3.0
DATE 26-8-2003
Error ContentrEncoding -> ContentEncoding
same as version 2.9 but now we keep in sync with the documentation
versions that are 3.0 (0-..)

VERSION 2.9
DATE 3-7-2003
- Added maxOccurs="unbounded" to Content.SubGenre
- Added maxOccurs="unbounded" to Content.Task
- Added maxOccurs="unbounded" to Content.Modalities
- Added minOccurs="0" to Session.Description
- Added minOccurs="0" to Actor.Description
- Added minOccurs="0" to MediaFile.Description
- Added minOccurs="0" to WrittenResource.Description
- Added minOccurs="0" to Source.Description
- Added minOccurs="0" to Validation.Description
- Added minOccurs="0" to Access.Description

FollowUp attribute in CVTypeDef introduced

VERSION 2.8
DATE 23-6-2003

- Added content encoding to written resource
DATE 18-6-2003
- Added attribute group to cater for Session profiles
- Added encoding attribute to Content.Subject
- changed multiplicity restrictions for some elements

VERSION 2.7
DATE 17-6-2003
Changed structure of TimePositionType and CounterPositionType
to complex types. No information in attributes only in subelements

VERSION 2.6
DATE 30-5-2003
imdi:boolean had typo, MotherTongue & PrimaryLanguage type to imdi:boolean
corresponds to document version 3.02, in final version synchronise version numbers!

VERSION 2.5
DATE 21-5-2003
Added imdi:boolean type
Addd ResourceRef attribute to Source
corresponds to document version 3.02, in final version synchronise version numbers!

VERSION 2.4
DATE 13-5-2003
Added Lexicon resource
corresponds to document version 3.02, in final version synchronise version numbers!

VERSION: 2.3
DATE: 2003-04-16
Checked cardinalities
Only one Location (was unbounded)
At least one Description in (optional) References

DATE 2003-04-08.
Corresponds to document version 3.02, in final version synchronise version numbers!

VERSION: 2.2
DATE: 2003-03-06
-->
<xsd:schema targetNamespace="http://www.mpi.nl/IMDI/Schema/IMDI"
xmlns:imdi="http://www.mpi.nl/IMDI/Schema/IMDI" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" attributeFormDefault="unqualified" version="1">
  <xsd:element name="METATRANSCRIPT">
    <xsd:annotation>
      <xsd:documentation>The outer element with administrative data of all metadata description files.
Version 1.0 is based on Session description version 2.5 and Catalogue description 2.1 Version 3.0 is
based on Session description version 3.03, Catalogue description 2.1, Lexicon description
1.1</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType mixed="false">
      <xsd:sequence>
        <xsd:element name="History" type="xsd:string" minOccurs="0">
          <xsd:annotation>
            <xsd:documentation>Creation history of this metadata descriptionfile</xsd:documentation>
          </xsd:annotation>
        </xsd:element>
        <xsd:choice>
          <xsd:element name="Session" type="imdi:SessionType" maxOccurs="unbounded"/>
          <xsd:element name="Corpus" type="imdi:CorpusType" maxOccurs="unbounded"/>
          <xsd:element name="Catalogue" type="imdi:CatalogueType"/>

```
      </xsd:choice>
    </xsd:sequence>
    <xsd:attribute name="Profile" type="xsd:string" use="optional"/>
    <xsd:attribute name="Date" type="xsd:date" use="required"/>
    <xsd:attribute name="Originator" type="xsd:string" use="optional"/>
    <xsd:attribute name="Version" type="xsd:string" use="required"/>
    <xsd:attribute name="FormatId" type="xsd:string" use="required"/>
    <xsd:attribute name="History" type="xsd:anyURI" use="optional"/>
    <xsd:attribute name="Type" type="imdi:MetatranscriptType" use="required"/>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
 </xsd:element>
 <xsd:attributeGroup name="ProfileAttributes">
  <xsd:attribute name="XXX-Type" type="xsd:string" use="optional"/>
  <xsd:attribute name="XXX-Multiple" type="xsd:boolean" use="optional"/>
  <xsd:attribute name="XXX-Visible" type="xsd:boolean" use="optional"/>
  <xsd:attribute name="XXX-Tag" type="xsd:string" use="optional"/>
  <xsd:attribute name="XXX-HelpText" type="xsd:string" use="optional"/>
  <xsd:attribute name="XXX-FollowUpDepend" type="xsd:string" use="optional"/>
 </xsd:attributeGroup>
 <xsd:simpleType name="boolean">
  <xsd:annotation>
   <xsd:documentation>boolean or unknown / unspecified value</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
   <xsd:enumeration value="true"/>
   <xsd:enumeration value="false"/>
   <xsd:enumeration value="Unknown"/>
   <xsd:enumeration value="Unspecified"/>
  </xsd:restriction>
 </xsd:simpleType>
 <xsd:complexType name="NamedLinkType">
  <xsd:annotation>
   <xsd:documentation>link to other resource. Attribute name is for the benefit of
browsing</xsd:documentation>
  </xsd:annotation>
  <xsd:simpleContent>
   <xsd:extension base="xsd:anyURI">
    <xsd:attribute name="Name" type="xsd:string" use="required"/>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
   </xsd:extension>
  </xsd:simpleContent>
 </xsd:complexType>
 <xsd:complexType name="DescriptionType" block="extension" mixed="false">
  <xsd:annotation>
   <xsd:documentation>Human readable description in the form of a text with language id
specification and/or a link to a file with a description and language id specification. The name attribute
is to name the link (if present)</xsd:documentation>
  </xsd:annotation>
  <xsd:simpleContent>
   <xsd:extension base="xsd:string">
    <xsd:attribute name="LanguageId" type="imdi:LanguageIdType" use="optional"/>
    <xsd:attribute name="Name" type="xsd:string" use="optional"/>
    <xsd:attribute name="Link" type="xsd:anyURI" use="optional"/>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
   </xsd:extension>
  </xsd:simpleContent>
 </xsd:complexType>
 <xsd:complexType name="ContactType">
  <xsd:annotation>
```

```
      <xsd:documentation>Contact information for this data</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Name" minOccurs="0">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Address" minOccurs="0">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Email" minOccurs="0">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Organisation" minOccurs="0">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
  <xsd:complexType name="LocationType">
    <xsd:annotation>
      <xsd:documentation>Information on creation location for this data</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Continent">
        <xsd:annotation>
          <xsd:documentation>The continent where the session/corpus was recorded or
originated</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:Vocabulary"/>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Country">
        <xsd:annotation>
```

```
        <xsd:documentation>The country where the session/corpus was recorded or
originated</xsd:documentation>
       </xsd:annotation>
       <xsd:complexType>
        <xsd:simpleContent>
         <xsd:extension base="imdi:Vocabulary"/>
        </xsd:simpleContent>
       </xsd:complexType>
      </xsd:element>
      <xsd:element name="Region" minOccurs="0" maxOccurs="unbounded">
       <xsd:annotation>
        <xsd:documentation>The region or sub-region where the session/corpus was recorded or
originated</xsd:documentation>
       </xsd:annotation>
       <xsd:complexType>
        <xsd:simpleContent>
         <xsd:extension base="xsd:string">
          <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
         </xsd:extension>
        </xsd:simpleContent>
       </xsd:complexType>
      </xsd:element>
      <xsd:element name="Address" minOccurs="0">
       <xsd:annotation>
        <xsd:documentation>The address where the session/corpus was recorded or
originated</xsd:documentation>
       </xsd:annotation>
       <xsd:complexType>
        <xsd:simpleContent>
         <xsd:extension base="xsd:string">
          <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
         </xsd:extension>
        </xsd:simpleContent>
       </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
   </xsd:complexType>
   <xsd:element name="Keys">
    <xsd:annotation>
     <xsd:documentation>List of a number of key name value pairs. Should be used to add information
that is not covered by other metadata elements at this level</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType mixed="false">
     <xsd:sequence>
      <xsd:element name="Key" minOccurs="0" maxOccurs="unbounded">
       <xsd:complexType>
        <xsd:simpleContent>
         <xsd:extension base="imdi:Vocabulary">
          <xsd:attribute name="Name" type="xsd:string" use="required"/>
         </xsd:extension>
        </xsd:simpleContent>
       </xsd:complexType>
      </xsd:element>
     </xsd:sequence>
     <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:complexType>
   </xsd:element>
   <xsd:element name="Languages">
    <xsd:annotation>
```

```
      <xsd:documentation>Groups information about the languages used in the
session</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType mixed="false">
      <xsd:sequence>
        <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
          <xsd:annotation>
            <xsd:documentation>Description for the list of languages spoken by this
participant</xsd:documentation>
          </xsd:annotation>
        </xsd:element>
        <xsd:element ref="imdi:Language" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Language">
    <xsd:annotation>
      <xsd:documentation>An element from a set of languages used in the
session</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType mixed="false">
      <xsd:sequence>
        <xsd:element name="Id">
          <xsd:annotation>
            <xsd:documentation>Unique code to identify a language</xsd:documentation>
          </xsd:annotation>
          <xsd:complexType>
            <xsd:simpleContent>
              <xsd:extension base="imdi:LanguageIdType">
                <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
              </xsd:extension>
            </xsd:simpleContent>
          </xsd:complexType>
        </xsd:element>
        <xsd:element name="Name" type="imdi:Vocabulary" maxOccurs="unbounded">
          <xsd:annotation>
            <xsd:documentation>A list of human understandable names to identify a
language</xsd:documentation>
          </xsd:annotation>
        </xsd:element>
        <xsd:element name="MotherTongue" minOccurs="0">
          <xsd:annotation>
            <xsd:documentation>Is it the speakers mother tongue. Only applicable if used in the context
of a speakers language</xsd:documentation>
          </xsd:annotation>
          <xsd:complexType>
            <xsd:simpleContent>
              <xsd:extension base="imdi:boolean">
                <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
              </xsd:extension>
            </xsd:simpleContent>
          </xsd:complexType>
        </xsd:element>
        <xsd:element name="PrimaryLanguage" minOccurs="0">
          <xsd:annotation>
            <xsd:documentation>Is it the speakers primary language. Only applicable if used in the
context of a speakers language</xsd:documentation>
          </xsd:annotation>
```

```
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:boolean">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Dominant" minOccurs="0">
        <xsd:annotation>
          <xsd:documentation>Is it the most frequently used language in the document. Only applicable
if used in the context of the resource's language</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:boolean">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="SourceLanguage" minOccurs="0">
        <xsd:annotation>
          <xsd:documentation>Direction of translation. Only applicable in case it is the context of a
lexicon resource</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:boolean">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="TargetLanguage" minOccurs="0">
        <xsd:annotation>
          <xsd:documentation>Direction of translation. Only applicable in case it is the context of a
lexicon resource</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:boolean">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
        <xsd:annotation>
          <xsd:documentation>Description for this particular language</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
    </xsd:sequence>
    <xsd:attribute name="ResourceRef" type="xsd:string" use="optional"/>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
 </xsd:element>
 <xsd:simpleType name="LanguageIdType">
  <xsd:annotation>
```

```
      <xsd:documentation>Language identification either ISO or SIL</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="(ISO639(-1|-2)?:.*)?"/>
      <xsd:pattern value="(RFC3066:.*)?"/>
      <xsd:pattern value="(RFC1766:.*)?"/>
      <xsd:pattern value="(SIL:.*)?"/>
      <xsd:pattern value="Unknown"/>
      <xsd:pattern value="Unspecified"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="AccessType">
    <xsd:annotation>
      <xsd:documentation>Groups information about access rights for this data</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Availability" type="imdi:Vocabulary">
        <xsd:annotation>
          <xsd:documentation>Availability of the data</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
      <xsd:element name="Date">
        <xsd:annotation>
          <xsd:documentation>Date when access rights were evaluated</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:DateType"/>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Owner">
        <xsd:annotation>
          <xsd:documentation>Name of owner resource</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Publisher">
        <xsd:annotation>
          <xsd:documentation>Publisher responsible for distribution of this data</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Contact" type="imdi:ContactType"/>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
```

```
  </xsd:complexType>
  <xsd:complexType name="ExternalResourceReferenceType">
    <xsd:annotation>
      <xsd:documentation>Resource is preferably a metadata resource. In the case of a well-defined merged metadata/content format such as TEI or legacy resources for which no further metadata is available it is the resource itself. If the external resource is an IMDI session with written resources Type &amp;amp; SubType will be the same as the Type &amp;amp; SubType of the primary written resource in that session. If it is a session with IMDI multi-media resources the Type of the Media File will
        designate it. SubType is used only for written resources. Non-IMDI metadata resource types need to be mapped to IMDI types</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Type">
        <xsd:annotation>
          <xsd:documentation>The type of the external (metadata) resource</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:Vocabulary"/>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="SubType" minOccurs="0">
        <xsd:annotation>
          <xsd:documentation>The sub type of the external (metadata) resource. Only used in case its metadata for a written resource</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:Vocabulary"/>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Format">
        <xsd:annotation>
          <xsd:documentation>The metadata format</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:Vocabulary"/>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Link" type="xsd:anyURI">
        <xsd:annotation>
          <xsd:documentation>The URL of the external metadata record</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
  <xsd:complexType name="ProjectType">
    <xsd:annotation>
      <xsd:documentation>Project Information</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Name">
        <xsd:annotation>
          <xsd:documentation>A short name or abbreviation for the project</xsd:documentation>
        </xsd:annotation>
```

```
      <xsd:complexType>
       <xsd:simpleContent>
        <xsd:extension base="xsd:string">
         <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
        </xsd:extension>
       </xsd:simpleContent>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="Title">
      <xsd:annotation>
       <xsd:documentation>The full title of the project</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
       <xsd:simpleContent>
        <xsd:extension base="xsd:string">
         <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
        </xsd:extension>
       </xsd:simpleContent>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="Id">
      <xsd:annotation>
       <xsd:documentation>A unique identifier for the project</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
       <xsd:simpleContent>
        <xsd:extension base="xsd:string">
         <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
        </xsd:extension>
       </xsd:simpleContent>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="Contact" type="imdi:ContactType">
      <xsd:annotation>
       <xsd:documentation>Contact information for this project</xsd:documentation>
      </xsd:annotation>
     </xsd:element>
     <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
      <xsd:annotation>
       <xsd:documentation>Description for this project</xsd:documentation>
      </xsd:annotation>
     </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
   </xsd:complexType>
   <xsd:element name="CounterPosition">
    <xsd:annotation>
     <xsd:documentation>Position (start (+end) ) on a oldfashioned tape without time
indication</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType mixed="false">
     <xsd:sequence>
      <xsd:element name="Start" type="imdi:IntegerType"/>
      <xsd:element name="End" minOccurs="0" type="imdi:IntegerType"/>
     </xsd:sequence>
     <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:complexType>
   </xsd:element>
   <xsd:element name="TimePosition">
```

```
  <xsd:annotation>
    <xsd:documentation>Position in a media file or modern tape</xsd:documentation>
  </xsd:annotation>
  <xsd:complexType mixed="false">
    <xsd:sequence>
      <xsd:element name="Start">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:TimePositionType">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="End" minOccurs="0">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:TimePositionType">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
</xsd:element>
<xsd:simpleType name="emptyString">
  <xsd:annotation>
    <xsd:documentation>empty string definition</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:maxLength value="0"/>
  </xsd:restriction>
</xsd:simpleType>
<xsd:simpleType name="EmptyType">
  <xsd:annotation>
    <xsd:documentation>empty type definition</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="Unknown"/>
    <xsd:enumeration value="Unspecified"/>
  </xsd:restriction>
</xsd:simpleType>
<xsd:simpleType name="TimePositionType">
  <xsd:annotation>
    <xsd:documentation>Time position in the hh:mm:ss:ff format</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:pattern value="[0-9][0-9]:[0-9][0-9]:[0-9][0-9]:?[0-9]*|Unknown|Unspecified"/>
  </xsd:restriction>
</xsd:simpleType>
<!--
<xsd:complexType name="DateType">
  <xsd:simpleContent>
    <xsd:extension base="imdi:DateValueType">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>
```

```
  <xsd:simpleType name="IntegerType">
    <xsd:annotation>
      <xsd:documentation>integer + Unspecified and Unknown</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="[0-9]*|Unknown|Unspecified"/>
    </xsd:restriction>
  </xsd:simpleType>
-->
  <xsd:complexType name="IntegerType">
    <xsd:simpleContent>
      <xsd:extension base="imdi:imdiIntegerType">
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:extension>
    </xsd:simpleContent>
  </xsd:complexType>
  <xsd:simpleType name="imdiIntegerType">
    <xsd:annotation>
      <xsd:documentation>integer + Unspecified and Unknown</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="[0-9]*|Unknown|Unspecified"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:simpleType name="VocabularyType">
    <xsd:annotation>
      <xsd:documentation>specifies the four vocabulary types</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="ClosedVocabulary"/>
      <xsd:enumeration value="ClosedVocabularyList"/>
      <xsd:enumeration value="OpenVocabulary"/>
      <xsd:enumeration value="OpenVocabularyList"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:simpleType name="VocabularyRefType">
    <xsd:annotation>
      <xsd:documentation>Pointer to a vocabulary definition</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:anyURI"/>
  </xsd:simpleType>
  <xsd:complexType name="Vocabulary">
    <xsd:annotation>
      <xsd:documentation>value for an element/attribute that is a vocabulary</xsd:documentation>
    </xsd:annotation>
    <xsd:simpleContent>
      <xsd:extension base="imdi:CVSstring">
        <xsd:attribute name="Type" type="imdi:VocabularyType" default="OpenVocabulary"/>
        <xsd:attribute name="DefaultLink" type="imdi:VocabularyRefType" use="optional"/>
        <xsd:attribute name="Link" type="imdi:VocabularyRefType" use="optional"/>
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:extension>
    </xsd:simpleContent>
  </xsd:complexType>
  <xsd:complexType name="ValidationType">
    <xsd:annotation>
      <xsd:documentation>The validation used for the resource</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Type">
```

```
      <xsd:annotation>
        <xsd:documentation>CV: content, type, manual, automatic, semi-
automatic</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="imdi:Vocabulary"/>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Methodology" type="imdi:Vocabulary"/>
    <xsd:element name="Level" minOccurs="0" type="imdi:IntegerType">
      <xsd:annotation>
        <xsd:documentation>Percentage of resource done</xsd:documentation>
      </xsd:annotation>
    </xsd:element>
    <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded"/>
  </xsd:sequence>
  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
 </xsd:complexType>
 <xsd:complexType name="AgeType">
  <xsd:annotation>
    <xsd:documentation>Specifies age of a person with differerent counting
methods</xsd:documentation>
  </xsd:annotation>
  <xsd:simpleContent>
    <xsd:extension base="imdi:AgeValueType">
      <xsd:attribute name="AgeCountingMethod" type="imdi:AgeCountingMethodType"
default="SinceBirth"/>
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:extension>
  </xsd:simpleContent>
 </xsd:complexType>
 <xsd:complexType name="AgeRangeType">
  <xsd:annotation>
    <xsd:documentation>Specifies age of a person in the form of a range</xsd:documentation>
  </xsd:annotation>
  <xsd:simpleContent>
    <xsd:extension base="imdi:AgeRangeValueType">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:extension>
  </xsd:simpleContent>
 </xsd:complexType>
 <xsd:simpleType name="AgeCountingMethodType">
  <xsd:annotation>
    <xsd:documentation>The counting method</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="SinceConception"/>
    <xsd:enumeration value="SinceBirth"/>
  </xsd:restriction>
 </xsd:simpleType>
 <xsd:simpleType name="AgeValueType">
  <xsd:annotation>
    <xsd:documentation>The age of a person</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:pattern value="([0-9]+)*(;[0-9]+)*(.[0-9]+)*|Unknown|Unspecified"/>
  </xsd:restriction>
```

```
  </xsd:simpleType>
  <xsd:simpleType name="AgeRangeValueType">
    <xsd:annotation>
      <xsd:documentation>The age of a person given as a range</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="([0-9]+)?(;[0-9]+)?(.[0-9]+)?(/([0-9]+)?(;[0-9]+)?(.[0-
9]+)?)?|Unknown|Unspecified"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:simpleType name="CVSstring">
    <xsd:annotation>
      <xsd:documentation>Comma seperated string</xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="[^,]*(,[^,]+)*"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="MDGroupType">
    <xsd:annotation>
      <xsd:documentation>Type for group of metadata pertaining to a session</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Location" type="imdi:LocationType">
        <xsd:annotation>
          <xsd:documentation>Groups information about the location where the session was
created</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
      <xsd:element name="Project" type="imdi:ProjectType" maxOccurs="unbounded">
        <xsd:annotation>
          <xsd:documentation>Groups information about the project for which the session was (originally)
created</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
      <xsd:element ref="imdi:Keys"/>
      <xsd:element name="Content">
        <xsd:annotation>
          <xsd:documentation>Groups information about the content of the session. The content
description takes place in several (overlapping) dimensions</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="Genre" type="imdi:Vocabulary">
              <xsd:annotation>
                <xsd:documentation>Major genre classification</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="SubGenre" type="imdi:Vocabulary" minOccurs="0"
maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>Sub genre classification</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="Task" type="imdi:Vocabulary" minOccurs="0"
maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>List of he major tasks carried out in the session</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
```

```
        <xsd:element name="Modalities" type="imdi:Vocabulary" minOccurs="0"
maxOccurs="unbounded">
          <xsd:annotation>
            <xsd:documentation>List of modalities used in the session</xsd:documentation>
          </xsd:annotation>
        </xsd:element>
        <xsd:element name="Subject" minOccurs="0" maxOccurs="unbounded">
          <xsd:annotation>
            <xsd:documentation>Classifies the subject of the session. Uses preferably an existing
library classification scheme such as LCSH. The element has a scheme attribute that indicates what
scheme is used. Comments: The element can be repeated but the user should guarantee
consistency</xsd:documentation>
          </xsd:annotation>
          <xsd:complexType>
            <xsd:simpleContent>
              <xsd:extension base="imdi:Vocabulary">
                <xsd:attribute name="Encoding" type="xsd:string" use="optional"/>
              </xsd:extension>
            </xsd:simpleContent>
          </xsd:complexType>
        </xsd:element>
        <xsd:element name="CommunicationContext">
          <xsd:annotation>
            <xsd:documentation>This groups information concerning the context of
communication</xsd:documentation>
          </xsd:annotation>
          <xsd:complexType>
            <xsd:sequence>
              <xsd:element name="Interactivity" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>degree of interactivity</xsd:documentation>
                </xsd:annotation>
              </xsd:element>
              <xsd:element name="PlanningType" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>Degree of planning of the event</xsd:documentation>
                </xsd:annotation>
              </xsd:element>
              <xsd:element name="Involvement" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>Indicates in how far the researcher was involved in the linguistic
event</xsd:documentation>
                </xsd:annotation>
              </xsd:element>
              <xsd:element name="SocialContext" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>Indicates the social context the event took place
in</xsd:documentation>
                </xsd:annotation>
              </xsd:element>
              <xsd:element name="EventStructure" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>Indicates the structure of the communication
event</xsd:documentation>
                </xsd:annotation>
              </xsd:element>
              <xsd:element name="Channel" type="imdi:Vocabulary" minOccurs="0">
                <xsd:annotation>
                  <xsd:documentation>Indicates the channel of the communication</xsd:documentation>
                </xsd:annotation>
```

```
            </xsd:element>
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element ref="imdi:Languages"/>
      <xsd:element ref="imdi:Keys"/>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
        <xsd:annotation>
          <xsd:documentation>Description for the content of this session</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
</xsd:element>
<xsd:element name="Actors">
  <xsd:annotation>
    <xsd:documentation>Groups information about all actors in the session</xsd:documentation>
  </xsd:annotation>
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
        <xsd:annotation>
          <xsd:documentation>Description pertaining to all the actors together</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
      <xsd:element name="Actor" minOccurs="0" maxOccurs="unbounded">
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="Role" type="imdi:Vocabulary">
              <xsd:annotation>
                <xsd:documentation>Functional role of the actor e.g. consultant, contributor,
interviewer, researcher, publisher, collector, translator</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="Name" maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>Name of the actor as used by others in the
transcription</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="FullName">
              <xsd:annotation>
                <xsd:documentation>Official name of the actor</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
```

```
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Code">
            <xsd:annotation>
              <xsd:documentation>Short unique code to identify the participant as used in the
transcription</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:string">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="FamilySocialRole" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>The role of the participant in the session within the context of
informant</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element ref="imdi:Languages"/>
          <xsd:element name="EthnicGroup" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>List of ethnic groups of participant</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Age" type="imdi:AgeRangeType">
            <xsd:annotation>
              <xsd:documentation>Age or age range of the participant in CHAT format yy;mm.dd
(separated by an '/' in case of a range)</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="BirthDate" type="imdi:DateType"/>
          <xsd:element name="Sex" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>Sex of the participant</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Education">
            <xsd:annotation>
              <xsd:documentation>The education of the participant Can also be used to specify
litteracy</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:string">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Anonymized">
            <xsd:annotation>
              <xsd:documentation>Indicated if true names were used or that codes were
employed</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="imdi:boolean">
```

```
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="Contact" type="imdi:ContactType" minOccurs="0">
              <xsd:annotation>
                <xsd:documentation>Contact information of the actor</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element ref="imdi:Keys"/>
            <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>Description for this individual participant</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
          </xsd:sequence>
          <xsd:attribute name="ResourceRef" type="xsd:string" use="optional"/>
          <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
 </xsd:complexType>
 <xsd:complexType name="CorpusType">
  <xsd:annotation>
   <xsd:documentation>Type for a corpus that points to either other corpora or
sessions</xsd:documentation>
  </xsd:annotation>
  <xsd:sequence>
   <xsd:element name="Name">
    <xsd:annotation>
     <xsd:documentation>name of the (sub-)corpus</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType>
     <xsd:simpleContent>
      <xsd:extension base="xsd:string">
       <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:extension>
     </xsd:simpleContent>
    </xsd:complexType>
   </xsd:element>
   <xsd:element name="Title">
    <xsd:annotation>
     <xsd:documentation>Title for the (sub-)corpus</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType>
     <xsd:simpleContent>
      <xsd:extension base="xsd:string">
       <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:extension>
     </xsd:simpleContent>
    </xsd:complexType>
   </xsd:element>
   <xsd:element name="Description" type="imdi:DescriptionType" maxOccurs="unbounded"/>
```

```
      <xsd:element name="MDGroup" type="imdi:MDGroupType" minOccurs="0"/>
      <xsd:element name="CorpusLink" type="imdi:NamedLinkType" minOccurs="0"
maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="SearchService" type="xsd:anyURI" use="optional"/>
    <xsd:attribute name="CorpusStructureService" type="xsd:anyURI" use="optional"/>
    <xsd:attribute name="CatalogueLink" type="xsd:anyURI" use="optional"/>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
  <xsd:complexType name="CatalogueType">
    <xsd:annotation>
      <xsd:documentation>Type for group metadata pertaining to published
corpora</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Name">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Title">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Id" maxOccurs="unbounded">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Description" type="imdi:DescriptionType" maxOccurs="unbounded"/>
      <xsd:element name="DocumentLanguages">
        <xsd:annotation>
          <xsd:documentation>Groups information about the languages used for documentation of the
corpus</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType mixed="false">
          <xsd:sequence>
            <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>Description for the list of languages</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="Language" type="imdi:SimpleLanguageType" minOccurs="0"
maxOccurs="unbounded"/>
          </xsd:sequence>
          <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
```

```
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="SubjectLanguages">
      <xsd:annotation>
        <xsd:documentation>Groups information about the languages in the corpus that are subject of
analysis</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType mixed="false">
        <xsd:sequence>
          <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
            <xsd:annotation>
              <xsd:documentation>Description for the list of languages</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Language" type="imdi:SubjectLanguageType" minOccurs="0"
maxOccurs="unbounded"/>
        </xsd:sequence>
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Location" type="imdi:LocationType"/>
    <xsd:element name="ContentType">
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="imdi:Vocabulary"/>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Format">
      <xsd:complexType>
        <xsd:all>
          <xsd:element name="Text" type="imdi:Vocabulary" minOccurs="0"/>
          <xsd:element name="Audio" type="imdi:Vocabulary" minOccurs="0"/>
          <xsd:element name="Video" type="imdi:Vocabulary" minOccurs="0"/>
        </xsd:all>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Quality">
      <xsd:complexType>
        <xsd:all>
          <xsd:element name="Audio" type="imdi:QualityType" minOccurs="0"/>
          <xsd:element name="Video" type="imdi:QualityType" minOccurs="0"/>
        </xsd:all>
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="SmallestAnnotationUnit" type="imdi:Vocabulary"/>
    <xsd:element name="Applications" type="imdi:Vocabulary"/>
    <xsd:element name="Date" type="imdi:DateType"/>
    <xsd:element name="Project" type="imdi:ProjectType"/>
    <xsd:element name="Publisher">
      <xsd:annotation>
        <xsd:documentation>Publisher responsible for distribution of this data</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="xsd:string">
            <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:extension>
```

```
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Authors">
      <xsd:annotation>
        <xsd:documentation>Authors for the resources</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="imdi:CVSstring">
            <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:extension>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Size">
      <xsd:annotation>
        <xsd:documentation>Human readabusle string that indicates total size of
corpus</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="xsd:string">
            <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:extension>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="DistributionForm" type="imdi:Vocabulary"/>
    <xsd:element name="Access" type="imdi:AccessType"/>
    <xsd:element name="Pricing">
      <xsd:annotation>
        <xsd:documentation>Pricing info of the corpus</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="xsd:string">
            <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:extension>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
    <xsd:element ref="imdi:Keys"/>
  </xsd:sequence>
  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
 </xsd:complexType>
 <xsd:complexType name="SimpleLanguageType">
  <xsd:annotation>
    <xsd:documentation>Information on language name and id</xsd:documentation>
  </xsd:annotation>
  <xsd:sequence>
    <xsd:element name="Id">
      <xsd:annotation>
        <xsd:documentation>Unique code to identify a language</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="imdi:LanguageIdType">
            <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:extension>
```

```
      </xsd:simpleContent>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Name">
    <xsd:annotation>
      <xsd:documentation>The name of the language</xsd:documentation>
    </xsd:annotation>
    <xsd:complexType>
      <xsd:simpleContent>
        <xsd:extension base="imdi:Vocabulary"/>
      </xsd:simpleContent>
    </xsd:complexType>
  </xsd:element>
 </xsd:sequence>
 <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
</xsd:complexType>
<xsd:complexType name="SubjectLanguageType">
 <xsd:complexContent>
  <xsd:extension base="imdi:SimpleLanguageType">
   <xsd:sequence>
    <xsd:element name="Dominant" minOccurs="0">
     <xsd:annotation>
      <xsd:documentation>Indicates if language is dominant language</xsd:documentation>
     </xsd:annotation>
     <xsd:complexType>
      <xsd:simpleContent>
       <xsd:extension base="imdi:boolean">
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
       </xsd:extension>
      </xsd:simpleContent>
     </xsd:complexType>
    </xsd:element>
    <xsd:element name="SourceLanguage" minOccurs="0">
     <xsd:annotation>
      <xsd:documentation>Indicates if language is source language</xsd:documentation>
     </xsd:annotation>
     <xsd:complexType>
      <xsd:simpleContent>
       <xsd:extension base="imdi:boolean">
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
       </xsd:extension>
      </xsd:simpleContent>
     </xsd:complexType>
    </xsd:element>
    <xsd:element name="TargetLanguage" minOccurs="0">
     <xsd:annotation>
      <xsd:documentation>Indicates if language is target language</xsd:documentation>
     </xsd:annotation>
     <xsd:complexType>
      <xsd:simpleContent>
       <xsd:extension base="imdi:boolean">
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
       </xsd:extension>
      </xsd:simpleContent>
     </xsd:complexType>
    </xsd:element>
    <!--
   <xsd:element name="Dominant" type="imdi:boolean" minOccurs="0"/>
   <xsd:element name="SourceLanguage" type="imdi:boolean" minOccurs="0"/>
   <xsd:element name="TargetLanguage" type="imdi:boolean" minOccurs="0"/>
```

```
    -->
        <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
          <xsd:annotation>
            <xsd:documentation>Description of the language</xsd:documentation>
          </xsd:annotation>
        </xsd:element>
      </xsd:sequence>
    </xsd:extension>
  </xsd:complexContent>
 </xsd:complexType>
 <xsd:complexType name="DateType">
  <xsd:simpleContent>
    <xsd:extension base="imdi:DateValueType">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
    </xsd:extension>
  </xsd:simpleContent>
 </xsd:complexType>
 <xsd:simpleType name="DateValueType">
  <xsd:annotation>
    <xsd:documentation>Defines a date that can also be empty or Unknown or
Unspecified</xsd:documentation>
  </xsd:annotation>
  <xsd:union memberTypes="imdi:DateRangeValueType imdi:emptyString imdi:EmptyType"/>
 </xsd:simpleType>
 <xsd:simpleType name="DateRangeValueType">
  <xsd:annotation>
    <xsd:documentation>Defines a date range that can also be Unspecified or
Unknown</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:pattern value="([0-9]+)((-[0-9]+)(-[0-9]+)?)?(/([0-9]+)((-[0-9]+)(-[0-
9]+)?)?)?|Unknown|Unspecified"/>
  </xsd:restriction>
 </xsd:simpleType>
 <xsd:simpleType name="MetatranscriptType">
  <xsd:annotation>
    <xsd:documentation>Defines the different types of metatadata descriptions</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="SESSION"/>
    <xsd:enumeration value="SESSION.Profile"/>
    <xsd:enumeration value="LEXICON_RESOURCE_BUNDLE.Profile"/>
    <xsd:enumeration value="CATALOGUE"/>
    <xsd:enumeration value="CATALOGUE.Profile"/>
    <xsd:enumeration value="CORPUS"/>
    <xsd:enumeration value="CORPUS.Profile"/>
  </xsd:restriction>
 </xsd:simpleType>
 <xsd:simpleType name="QualityType">
  <xsd:annotation>
    <xsd:documentation>Quality specification scale [1-5] or Unknown or
Unspecified</xsd:documentation>
  </xsd:annotation>
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="1"/>
    <xsd:enumeration value="2"/>
    <xsd:enumeration value="3"/>
    <xsd:enumeration value="4"/>
    <xsd:enumeration value="5"/>
```

```
      <xsd:enumeration value="Unknown"/>
      <xsd:enumeration value="Unspecified"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="SessionType">
    <xsd:sequence>
      <xsd:element name="Name">
        <xsd:annotation>
          <xsd:documentation>A short name to identify the session</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Title">
        <xsd:annotation>
          <xsd:documentation>The complete title of the session without
abbrevations</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Date">
        <xsd:annotation>
          <xsd:documentation>The date when the primary data of the session was created (as a range)
in ISO8601 format (seperated by a '/')</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="imdi:DateRangeValueType">
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="ExternalResourceReference" type="imdi:ExternalResourceReferenceType"
minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded"/>
      <xsd:element name="MDGroup" type="imdi:MDGroupType"/>
      <xsd:element name="Resources">
        <xsd:annotation>
          <xsd:documentation>Groups information of language resources connected to the
session</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="MediaFile" minOccurs="0" maxOccurs="unbounded">
              <xsd:annotation>
                <xsd:documentation>Groups information about the media file</xsd:documentation>
              </xsd:annotation>
```

```
<xsd:complexType>
 <xsd:sequence>
  <xsd:element name="ResourceLink">
   <xsd:annotation>
    <xsd:documentation>URL to media file</xsd:documentation>
   </xsd:annotation>
   <xsd:complexType>
    <xsd:simpleContent>
     <xsd:extension base="xsd:anyURI">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
     </xsd:extension>
    </xsd:simpleContent>
   </xsd:complexType>
  </xsd:element>
  <xsd:element name="Type" type="imdi:Vocabulary">
   <xsd:annotation>
    <xsd:documentation>Major part of mime-type</xsd:documentation>
   </xsd:annotation>
  </xsd:element>
  <xsd:element name="Format" type="imdi:Vocabulary">
   <xsd:annotation>
    <xsd:documentation>Minor part of mime-type</xsd:documentation>
   </xsd:annotation>
  </xsd:element>
  <xsd:element name="Size">
   <xsd:annotation>
    <xsd:documentation>Size of media file</xsd:documentation>
   </xsd:annotation>
   <xsd:complexType>
    <xsd:simpleContent>
     <xsd:extension base="xsd:string">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
     </xsd:extension>
    </xsd:simpleContent>
   </xsd:complexType>
  </xsd:element>
  <xsd:element name="Quality">
   <xsd:annotation>
    <xsd:documentation>Quality of the recording scale [1-5]</xsd:documentation>
   </xsd:annotation>
   <xsd:complexType>
    <xsd:simpleContent>
     <xsd:extension base="imdi:QualityType">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
     </xsd:extension>
    </xsd:simpleContent>
   </xsd:complexType>
  </xsd:element>
  <xsd:element name="RecordingConditions">
   <xsd:annotation>
    <xsd:documentation>describes technical conditions of recording</xsd:documentation>
   </xsd:annotation>
   <xsd:complexType>
    <xsd:simpleContent>
     <xsd:extension base="xsd:string">
      <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
     </xsd:extension>
    </xsd:simpleContent>
   </xsd:complexType>
  </xsd:element>
```

```
            <xsd:element ref="imdi:TimePosition"/>
            <xsd:element name="Access" type="imdi:AccessType"/>
            <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded"/>
            <xsd:element ref="imdi:Keys"/>
           </xsd:sequence>
           <xsd:attribute name="ResourceId" type="xsd:string" use="optional"/>
           <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          </xsd:complexType>
         </xsd:element>
         <xsd:element name="WrittenResource" minOccurs="0" maxOccurs="unbounded">
          <xsd:annotation>
           <xsd:documentation>Groups information about a Written Resource</xsd:documentation>
          </xsd:annotation>
          <xsd:complexType>
           <xsd:sequence>
            <xsd:element name="ResourceLink">
             <xsd:annotation>
              <xsd:documentation>URL to file containing the
annotations/transcription</xsd:documentation>
             </xsd:annotation>
             <xsd:complexType>
              <xsd:simpleContent>
               <xsd:extension base="xsd:anyURI">
                <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
               </xsd:extension>
              </xsd:simpleContent>
             </xsd:complexType>
            </xsd:element>
            <xsd:element name="MediaResourceLink">
             <xsd:annotation>
              <xsd:documentation>URL to media file from which the annotations/transcriptions
originate </xsd:documentation>
             </xsd:annotation>
             <xsd:complexType>
              <xsd:simpleContent>
               <xsd:extension base="xsd:anyURI">
                <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
               </xsd:extension>
              </xsd:simpleContent>
             </xsd:complexType>
            </xsd:element>
            <xsd:element name="Date">
             <xsd:annotation>
              <xsd:documentation>Date when Written Resource was created</xsd:documentation>
             </xsd:annotation>
             <xsd:complexType>
              <xsd:simpleContent>
               <xsd:extension base="imdi:DateType"/>
              </xsd:simpleContent>
             </xsd:complexType>
            </xsd:element>
            <xsd:element name="Type" type="imdi:Vocabulary">
             <xsd:annotation>
              <xsd:documentation>The type of the WrittenResource</xsd:documentation>
             </xsd:annotation>
            </xsd:element>
            <xsd:element name="SubType" type="imdi:Vocabulary">
             <xsd:annotation>
              <xsd:documentation>The subtype of the WrittenResource</xsd:documentation>
```

```
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Format" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>File format used for Written Resource</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Size" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>The size of the Written Resource file. Integer value with addition of
M (mega) or K (kilo)</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Validation" type="imdi:ValidationType"/>
          <xsd:element name="Derivation" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>How this document relates to another
resource</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="CharacterEncoding">
            <xsd:annotation>
              <xsd:documentation>Character encoding used in the written
resource</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:string">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="ContentEncoding">
            <xsd:annotation>
              <xsd:documentation>Content encoding used in the written
resource</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:string">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="LanguageId" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>Language used in the resource</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Anonymized">
            <xsd:annotation>
              <xsd:documentation>Indicates if data has been anonymised. CV
boolean</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="imdi:boolean">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
```

```
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Access" type="imdi:AccessType"/>
          <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded"/>
          <xsd:element ref="imdi:Keys"/>
        </xsd:sequence>
        <xsd:attribute name="ResourceId" type="xsd:string" use="optional"/>
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="LexiconResource" minOccurs="0" maxOccurs="unbounded">
      <xsd:annotation>
        <xsd:documentation>Groups information only pertaining to a Lexical
resource</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:sequence>
          <xsd:element name="ResourceLink">
            <xsd:annotation>
              <xsd:documentation>URL to lexical resource</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:anyURI">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="MediaResourceLink">
            <xsd:annotation>
              <xsd:documentation>possible URL to media file connected to the lexical
resource</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:anyURI">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Date">
            <xsd:annotation>
              <xsd:documentation>Date when lexical resource was created</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="imdi:DateType"/>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Type" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>The type of the WrittenResource</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
```

```
            <xsd:element name="Format" type="imdi:Vocabulary">
              <xsd:annotation>
                <xsd:documentation>The format of the LexicalResource</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="CharacterEncoding">
              <xsd:annotation>
                <xsd:documentation>The character encoding of the
LexicalResource</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="SchemaRef">
              <xsd:annotation>
                <xsd:documentation>A reference to a possible structure schema for the
LR</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:anyURI">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="Size">
              <xsd:annotation>
                <xsd:documentation>The size of the LexicalResource in bytes</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="NoHeadEntries" type="imdi:IntegerType">
              <xsd:annotation>
                <xsd:documentation>The number of head entries of the
LexicalResource</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="NoSubEntries" type="imdi:IntegerType">
              <xsd:annotation>
                <xsd:documentation>The number of sub entries of the
LexicalResource</xsd:documentation>
              </xsd:annotation>
            </xsd:element>
            <xsd:element name="LexicalEntry" maxOccurs="unbounded">
              <xsd:complexType>
                <xsd:sequence>
                  <xsd:element name="HeadWordType" type="imdi:Vocabulary">
                    <xsd:annotation>
```

```
                    <xsd:documentation>OCV: Sentence, Phrase, Wordform, Lemma,
...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Orthography" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: HyphenatedSpelling, SyllabifiedSpelling,
...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Morphology" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: Stem,StemALlomorphy, Segmentation,
...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="MorphoSyntax" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: POS, Inflexion, Countability, ...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Syntax" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: Complementation, Alternation, Modification,
...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Phonology" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: Transcription, IPA Transcription, CV pattern,
...</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Semantics" type="imdi:Vocabulary">
                  <xsd:annotation>
                    <xsd:documentation>OCV: Sense dstinction</xsd:documentation>
                  </xsd:annotation>
                </xsd:element>
                <xsd:element name="Etymology" type="imdi:Vocabulary"/>
                <xsd:element name="Usage" type="imdi:Vocabulary"/>
                <xsd:element name="Frequency">
                  <xsd:complexType>
                    <xsd:simpleContent>
                      <xsd:extension base="xsd:string">
                        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                      </xsd:extension>
                    </xsd:simpleContent>
                  </xsd:complexType>
                </xsd:element>
              </xsd:sequence>
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="MetaLanguages">
            <xsd:annotation>
              <xsd:documentation>A block to describe the languages that are used to define terms, to
describe meaning</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:sequence>
```

```
                <xsd:element name="Language" type="imdi:Vocabulary"/>
                <xsd:element name="Description" type="imdi:DescriptionType"
maxOccurs="unbounded"/>
              </xsd:sequence>
              <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Access" type="imdi:AccessType"/>
          <xsd:element name="Description" type="imdi:DescriptionType"
maxOccurs="unbounded"/>
          <xsd:element ref="imdi:Keys"/>
        </xsd:sequence>
        <xsd:attribute name="ResourceId" type="xsd:string" use="optional"/>
        <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Source" minOccurs="0" maxOccurs="unbounded">
      <xsd:annotation>
        <xsd:documentation>Groups information about the source; e.g. media-carrier, book,
newspaper archive etc.</xsd:documentation>
      </xsd:annotation>
      <xsd:complexType>
        <xsd:sequence>
          <xsd:element name="Id">
            <xsd:annotation>
              <xsd:documentation>Short unique code to identify source</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="xsd:string">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element name="Format" type="imdi:Vocabulary">
            <xsd:annotation>
              <xsd:documentation>Physical storage format</xsd:documentation>
            </xsd:annotation>
          </xsd:element>
          <xsd:element name="Quality">
            <xsd:annotation>
              <xsd:documentation>Quality of recorded data. Scale [1-5]</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType>
              <xsd:simpleContent>
                <xsd:extension base="imdi:QualityType">
                  <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                </xsd:extension>
              </xsd:simpleContent>
            </xsd:complexType>
          </xsd:element>
          <xsd:element ref="imdi:CounterPosition" minOccurs="0"/>
          <xsd:element ref="imdi:TimePosition" minOccurs="0"/>
          <xsd:element name="Access" type="imdi:AccessType"/>
          <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
            <xsd:annotation>
              <xsd:documentation>Description for this source</xsd:documentation>
            </xsd:annotation>
```

```
            </xsd:element>
            <xsd:element ref="imdi:Keys"/>
          </xsd:sequence>
          <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
          <xsd:attribute name="ResourceRefs" type="xsd:string" use="optional"/>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Anonyms" minOccurs="0">
        <xsd:annotation>
          <xsd:documentation>Groups data about name conversions for persons who are
anonymised</xsd:documentation>
        </xsd:annotation>
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="ResourceLink">
              <xsd:annotation>
                <xsd:documentation>URL to information to convert pseudo named to real-
names</xsd:documentation>
              </xsd:annotation>
              <xsd:complexType>
                <xsd:simpleContent>
                  <xsd:extension base="xsd:anyURI">
                    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
                  </xsd:extension>
                </xsd:simpleContent>
              </xsd:complexType>
            </xsd:element>
            <xsd:element name="Access" type="imdi:AccessType"/>
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
</xsd:element>
<xsd:element name="References" minOccurs="0">
  <xsd:annotation>
    <xsd:documentation>Groups information about external documentation associated with this
session</xsd:documentation>
  </xsd:annotation>
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="Description" type="imdi:DescriptionType" minOccurs="0"
maxOccurs="unbounded">
        <xsd:annotation>
          <xsd:documentation>Every description is a reference</xsd:documentation>
        </xsd:annotation>
      </xsd:element>
    </xsd:sequence>
    <xsd:attributeGroup ref="imdi:ProfileAttributes"/>
  </xsd:complexType>
</xsd:element>
</xsd:sequence>
<xsd:attributeGroup ref="imdi:ProfileAttributes"/>
</xsd:complexType>
<xsd:complexType name="VocabularyDefType" mixed="false">
  <xsd:annotation>
    <xsd:documentation>The definition of a vocabulary. Attributes: Date of creattion, Link to origin.
Contails a Description be element to descr+++ ibe the domain of the vocabulary and a (unspecified)
number of value enries</xsd:documentation>
```

```
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Description" type="imdi:DescriptionType" maxOccurs="unbounded"/>
      <xsd:element name="Entry" maxOccurs="unbounded">
        <xsd:complexType>
          <xsd:simpleContent>
            <xsd:extension base="xsd:string">
              <xsd:attribute name="Tag" type="xsd:string" use="optional"/>
              <xsd:attribute name="Value" type="xsd:string" use="required"/>
            </xsd:extension>
          </xsd:simpleContent>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
    <xsd:attribute name="Name" type="xsd:string" use="required"/>
    <xsd:attribute name="Date" type="xsd:date" use="required"/>
    <xsd:attribute name="Tag" type="xsd:date" use="optional"/>
    <xsd:attribute name="Link" type="imdi:VocabularyRefType" use="required"/>
  </xsd:complexType>
  <xsd:element name="VocabularyDef" type="imdi:VocabularyDefType">
    <xsd:annotation>
      <xsd:documentation>Instantiation of a VocabularyDefType</xsd:documentation>
    </xsd:annotation>
  </xsd:element>
</xsd:schema>
```

Appendix E

Special Report

Distributed Access Management

Special Report
Distributed Access Management

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: Special Report

Date: 11.1.2006

Updated: 5.2.2007

Content

1. Introduction

This report is meant to summarize all the basics of the core of the DAM-LR project. Some of the points (URIDs) have already been agreed and are part of the "Definitions" deliverable. Others have to be discussed carefully with all partners involved at our coming Lund meeting in January 2006. All points addressed here and that have not yet been decided need this to become part of the official definitions document. This will allow us to develop missing software components. Of course, it must be possible to review such decisions later based on the experiences in the project, but we have to be very careful with revisions to not risk the success of the project.
In this sense this document has to be seen as essential and every partner team has to discuss the mentioned items intensively to be sure that we are on the right way. It should be mentioned that the essentials were presented at the DELAMAN meeting in Austin to get feedback from other archives not involved in DAM-LR. Actually, all suggestions were basically endorsed. The most intensly discussed points were the questions

- what a federation exactly is and
- which user attributes should be exchanged within a Shibboleth setup (see below).

We will not discuss the issue of what a federation is in this document in any larger scope. However it is obvious that this also has to be discussed. Let us not forget that DAM-LR has to be seen as a test case for a larger research infrastructure at European level and beyond. For DAM-LR we have to sort out the question, but it should be obvious that all points mentioned in this document have to be part of a federation declaration. All partners have to agree on the points mentioned in this note. Beyond these details, however, there has to be something like "trust", which is very difficult if not impossible to formalize. We suggest that David Nathan will distribute his note on "federations" as soon as possible.

2. PKI System

Basis 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[3] 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

The partners agreed on a number of issues here already. This is just to summarize the discussion. Details are described in the appendixes.

- 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 issued by the CNRI[4] 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 its handle postfix. This requires, however, that handle requests crossing the local boundaries

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

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 include any 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 fulfil 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. Authentication

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

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[5] as the basis for local authentication is recommended especially when it can be integrated with an already existing LDAP that is used to manage user identities of larger parent organisations.
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.

Therefore, the MPI will step over to Open LDAP for authentication for its own user management, which will include both 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 any formal affiliation. The DAM-LR core solution will rely on LDAP, all partners that will choose 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	(code for) hosting institute, that administrates the primary account for the user and where he can be authenticated (Shibboleth)
email address	email address of the user
status	status of a user in the institution, used at the discretion of archive managers
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 user identifier string within the federation 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

[5] 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.

tells us that it is always good to have a unique identifier in addition)

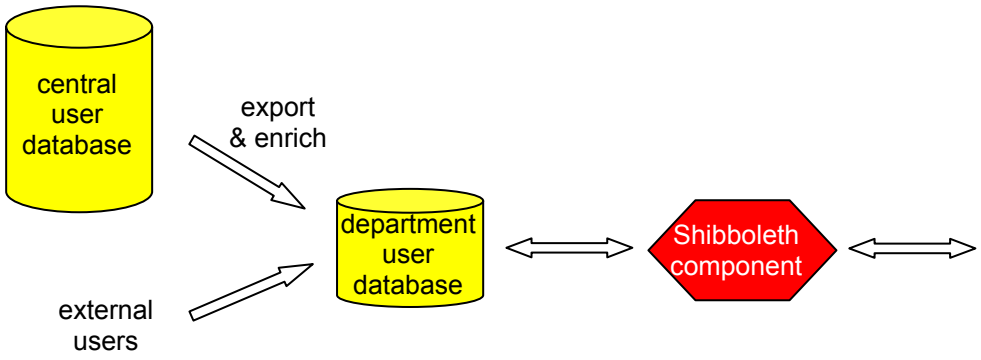
These attributes seem to be the minimal set and are largely overlapping with the specifications from EDU-Person and RFC 2798[6], others such as a host introducing an external guest as a collaborator or so may be necessary but are not yet fully identified. The MPI certainly will store internally aspects such as department, start of employment, end of employment and misbehaviour 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 misbehaviour 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 misbehaviour. Of course, we cannot prevent completely that the same person will register again under 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 there may be two problems:

it could be difficult to be home institute 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. We will split the discussion in 6 topics.

5.1 General Aspects

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

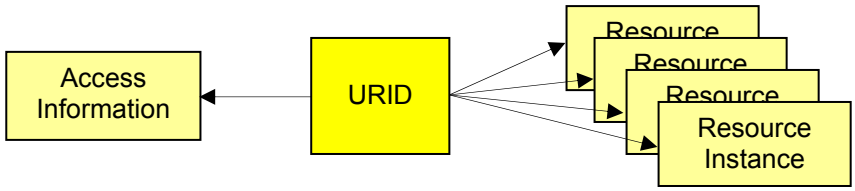
single identity to be achieved by authentication only by the user's home institute and accepting the connected attributes whenever he wants to access a resource in a federation archive
single sign-on once the user is identified he/she has transparent access to all resources he/she is permitted to access in all archives
one basket idea the user must be able to see his/her set of accessible resources as

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

	his/her temporary working archive
replication option	the archives must accept each other in so far that they exchange metadata and resources from each other.

One of the basic agreements in DAM-LR is that specifying authorization rules is done by the originating institution. Since for each resource independent of the number of instances there will be only one URID[7] also maintained by the originating institute, it looks advantageous to merge the authorization information with the URID record.

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.



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 we can use to 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 side has to know about the registered users. This has to be considered when designing software, since the administrational load can become intractable.

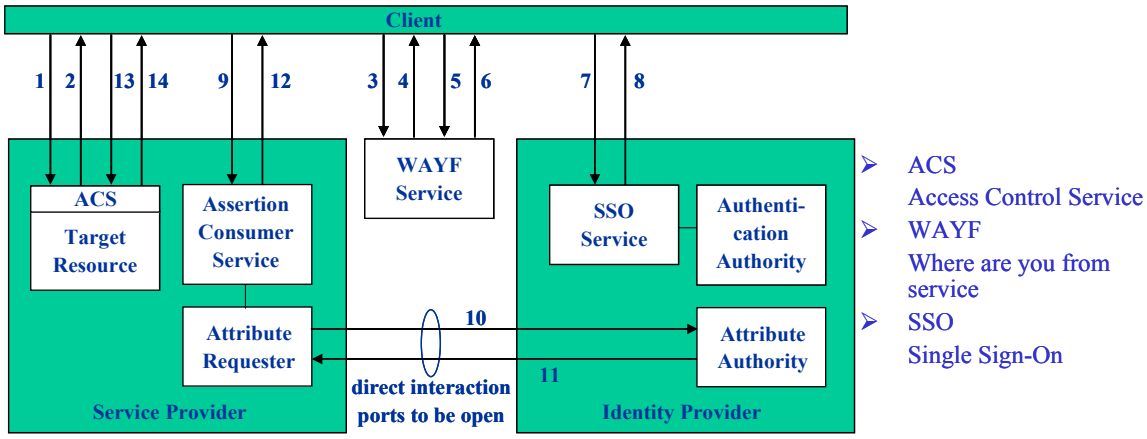
Another difference to the typical Shibboleth scenario (see below) is given by the fact that users will partly request access not to just one resource but to several (search across corpora, access to annotated media files etc). This also has to be taken into account.

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 an access scenario dominated by groups. Nevertheless, despite this scenario mismatch, 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. 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 Shibboleth puts responsibility for authentication at the home institute.

[7] 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.

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 authentication system makes sense. Shibboleth 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, but that is for later concern.



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.

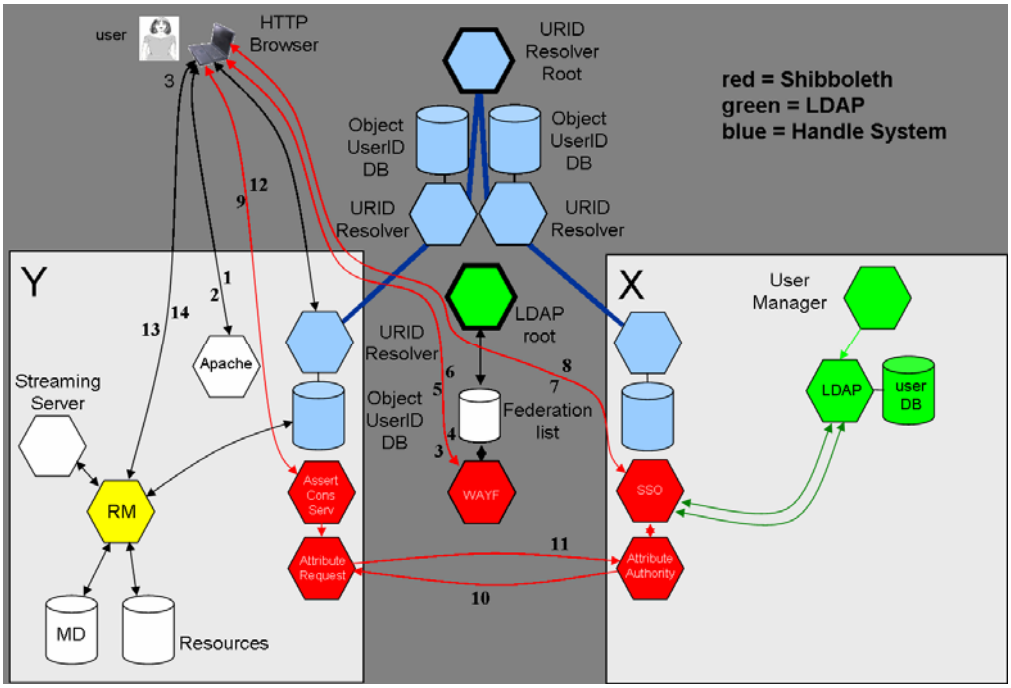
When analyzing the information flow with respect to repeated requests a few options seem to be possible:
The profile finally can contain all necessary information about a user such as user attributes, and session number. When the user wants to access another resource (1) all information is available 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. This interaction may be precluded by 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 this 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.



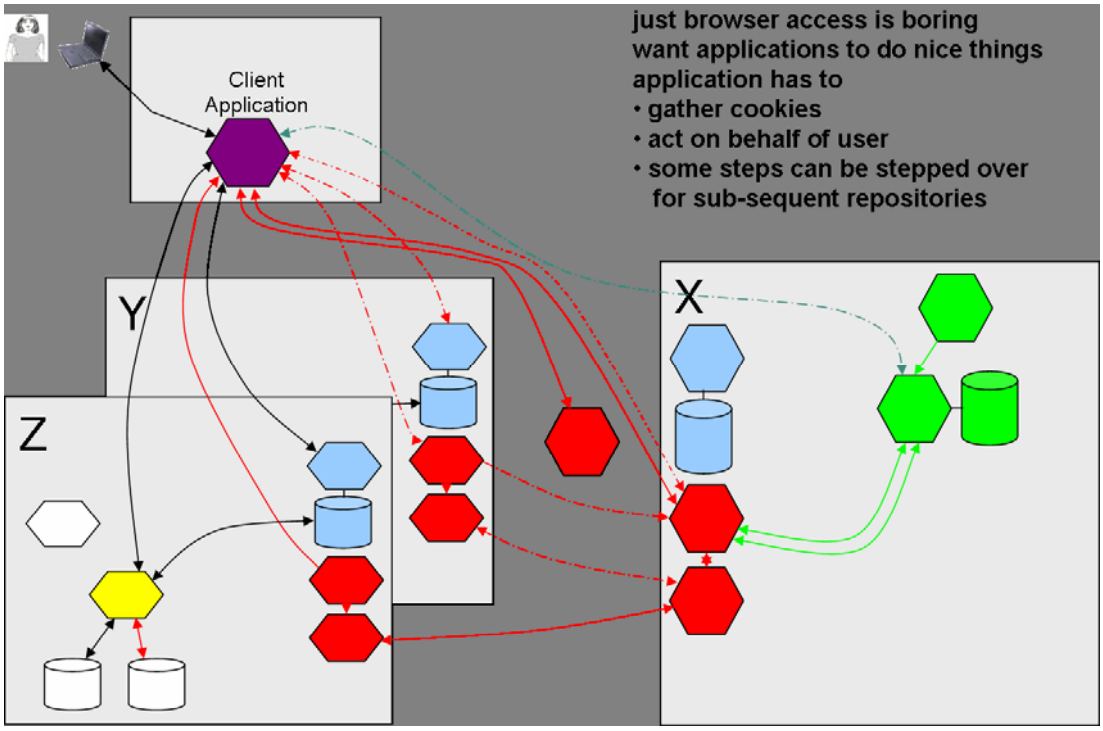
Shibboleth comes with a default resource manager that is integrated with the apache webserver
It is expected that this implementation is sufficient for our purposes, trying to develop a resource manager ourselves would not be cost effective.

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 operation may be requested on 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.

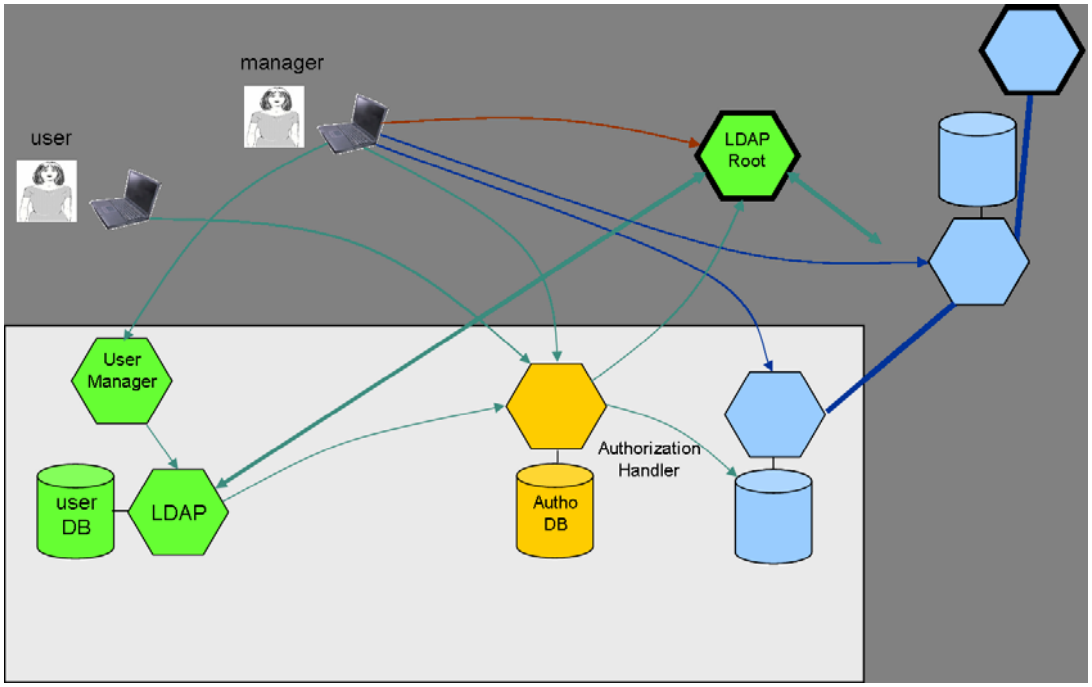
The application has now to act on behalf of the user if different sites are involved. Authentication will be requested by every site that has to be contacted, but the user should type it in just once. The application will have to mimic a web-browser's functions since using a web-browser to access multiple distributed resources protected by Shibboleth is been known to support this scenario.

So the applications have to be extended to support such scenarios. In a basket like situation it may make sense to do the authentication 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 institute. 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 institute and those who are accepted as guests.

If a user from a partner institute wants to access a resource, also a user record needs to be created in the LDAP in order that the correct user identifier can be used in the authorization records. A difference however is that now we can be certain of the users identity if we protect the request form with Shibboleth. Other possibilities to achieve this, can be to bind all LDAPs of the partners together in a joiny domain, but this would require a central authority and makes the use of LDAP as a authentication module mandatory.

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 can ascertain the claimed identity of the requester as explained above. 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 records on a high level.

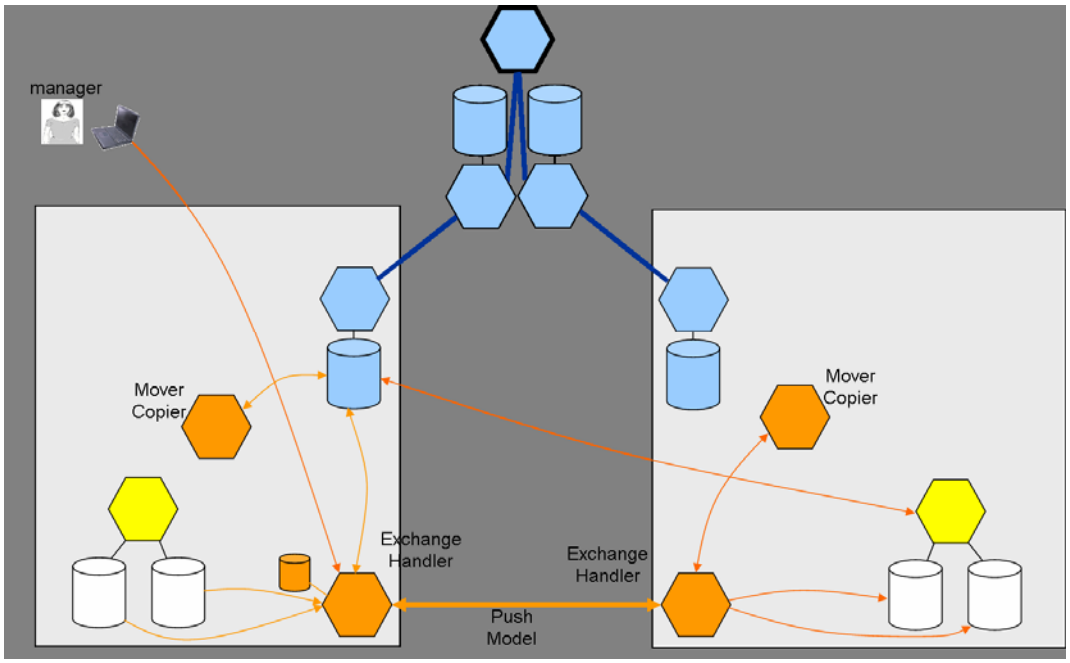
The Authorization DB also contains per sub-corpus specifications about processes such as whether the depositor has to be asked first, whether the person has to sign a declaration etc. When the user has fulfilled all required steps the general record will be expanded into corresponding formal records for the HTTP server and into URID access records[8] for dessimination 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. For local changes a mover/copier component has to be developed that updates the URID database for all modifications of physical paths of archive objects..

In the case that data will be exchanged between repositories[9] 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.

[8] 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.
[9] 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 regular scrutiny to see if they are still valid

6.1 Software Components and Certificates

The following software off-the-shelf components will be used within the prototypical system (Italics mean the component is required by DAM-LR for interoperability):

- 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 authentication and the exchange of user information*
- Apache as the 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:

- As a Shibboleth resource manager we will use the default shibd module for the Apache web server.
- A resource request form or web-application that when protected by Shibboleth will ascertain the provided user credentials. Other shared LDAP domain based solutions are also possible
- 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 is currently testing this component)

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 (those already agreed are in italics)

- *the Handle System will be used*
- *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 have to agree on a number of exchangeable user attributes in January
- the partners agree to carry out user management that will have relevance for DAM-LR in a careful and trustful way
- the partners have to 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

- Defining access rights is done by the originating[10] 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 Shibboleth
- 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 DAM-LR project; partners are free to develop their own component, but have to adhere to reliability requirements and carry out careful tests

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

- A prototypical web form or web-application for resource 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

Appendix F

Draft user agreement for the federation

DAM-LR 25-8-06
DN

User assents prior to access:

- primary purpose is to protect intellectual property and reduce risk of misuse; will express core ideas accepted by partners
- as a “delivery and agreement page”, it also provides opportunities for other processes to be inserted, e.g. user satisfaction survey etc, changes of conditions, news and updates, messages to users regarding e.g. registration or other access issues
- may point to web page of conditions signed up to at registration, and indicate (or link to) privacy and data protection information
- preferable to have a consistent formulation for all resources (i.e. not varying by resource - resource specific stuff should be separated and dealt with as a prior step to gain access path)
- question: what is the relation to other parts of authorisation process (Shibboleth etc); would it precede (e.g. to save messages/transactions), complicate, or perhaps fit right into one part of the authorisation process?

Access Agreement - Draft 2 (Aug 06)

In receiving electronic data ("the materials") from [repository], I acknowledge the following:

1. I am the individual named _____
2. "The materials" refers to any part of the data received, or any data derived from them.
3. The materials include metadata that describes the content of the materials and which may contain further conditions of use or information required for appropriate acknowledgement. This metadata may accompany other files within the materials or may be embedded in files within the materials.
4. I will use the materials only for private, not-for-profit research. Any other usage needs explicit written agreement from [repository].
5. I will not transfer the materials to any person, or do anything which could reasonably be understood to allow such transfer to take place.
6. If the materials are used in any manner in the preparation of any publication or other dissemination, I will acknowledge [repository] and the owner of the materials (click here to generate an example acknowledgement), and will, wherever possible, send an electronic copy of such publication to [repository].
7. I understand that [repository] may like to receive a copy of any new or revised resource created using the materials, in order to improve its collection.
8. I understand that language materials may contain sensitive information, and that I may not use them for any purpose which creates disparagement, disrespect, damage to reputation, or harm to any individual or group.
9. I understand that the materials are provided by [repository] as supplied, and that they are not guaranteed to fulfil any purpose or function, or to be compatible with any particular computing system or software. I release and indemnify [repository] from any consequences or damages resulting from my receipt or use of the materials.
10. I understand that [repository] records information about the access of these materials in accordance with the XXX laws of the United Kingdom and the European Union, and that this recorded information may identify me personally and may be passed to third parties (e.g. the owner of the materials) for purposes relevant to the management of the materials.
11. I have read this agreement and [repository’s] conditions of access ([web page basically replicating this stuff]) and I agree to be bound by them. I understand that [repository] may take legal action on behalf of itself or the owner of the materials if access conditions or any other rights held by [repository] or the owner are infringed.
12. [ANY ADDITIONAL CONDITIONS RELATED TO THE PARTICULAR MATERIALS TO BE EXPRESSED OR LINKED HERE]

Agree ❑ Disagree ❑ Name [auto filled in] Date [auto filled in]

Appendix G

SOAS Handle postfix system

SOAS HANDLE FORMAT

Resources are to be split up into four categories:

- Electronic archive objects
  - ELDP related (**E**)
  - Non-ELDP related (**N**)
- Others
  - Application binaries (**B**)
  - Physical objects (**P**)

Deposits, files and bundles get unique (PK) numeric IDs from the DB on accession / update; use these as part of the handle postfix

The Handles will be in the format:

```
[Category] / [Granularity] / [ID]
```

Example:

```
E/deposit/0123-4567-89AB-CDEF
N/bundle/FFAB-ABDA-CAED-230F
B/file/2009-1976-2508-2006
```

The numeric (id) part of the Handle postfix will use the same Hex-coding as the other partners use.

Appendix H

Access rights at the INL

Inventory of access rights for INL-LRs

INL DAM-LR workgroup
October 2005

1 Introduction

The Institute for Dutch Lexicology is one of the members of the DAM-LR[11] project. One of the tasks (work package 7) is to install a local solution for all four access management pillars[12] for the INL language resources (LRs). This document is the first step towards such a solution. It lists (the main categories of) INL LRs – our archive – and describes (our initial thoughts on) the access rights per resource (category).
The information in this document describes our specific situation, but some of it may be of relevance to other DAM-LR project members.

2 Access rights specification

A minimal set of access rights consists of read and write rights[13]. It is also useful to have entities to give these rights to: users and groups[14]. Abbreviations and a simple syntax are used to formalise permissions.
A summary of permission symbols used in this document:

U (User): individual users.
G (Group): groups of users
R (Read): read access
W (Write) write access

These symbols are used in the following syntax:

[11] Distributed Access Management for Language Resources. See http://www.mpi.nl/dam-lr
[12] Distributed Metadata, Unique Resource Identifiers, User and Group Management and Access Management
[13] At the time of writing, the INL does not yet support uploading of new data to our repository or changing of existing data. A system for managing and uploading of data to language archives (LAMUS) has been developed at (and in cooperation with) the Max Planck Institute, but not yet installed at the INL.
[14] This includes groups of groups.

Data User|Group Right

In case of more than one users or groups, use commas. Multiple access rights (e.g. read, write, execute) are grouped together (e.g. RWX).

Using this information, we can summarise "User Remco van Veenendaal has read access to the CGN corpus" as:

All CGN data: U:RemcoVanVeenendaal R

"User Remco van Veenendaal and the group of users who have a full licence for the CGN corpus have read and write access to the CGN corpus" is abbreviated to:

All CGN data: U:RemcoVanVeenendaal,G:CgnFullLicence RW

Please note that it would be preferable to make user Remco van Veenendaal member of the group CgnFullLicence.

3 Further thoughts on access rights

At some point in time (before accessing the language resources), users will have to sign a licence or otherwise request access. The user's name, password and access rights will then be added to some access management database, while the signed licence is stored somewhere safe. The procedure to obtain a licence and add users to the database is not part of this document[15]. Also, this document does not contain rules for deciding which user should be granted which level of access.

Users without a licence will have default access rights: they are considered members of the group Guests. The guest access rights are the lowest access level. All other access levels also have (inherit) guest access. At the lowest access level, users can access the metadata of all LRs[16]: all users have access to all metadata.

The access rights system described in this document does not distinguish between commercial, non-commercial users or any other type of user. Access has been granted (or not), that's all the system needs to know (see footnote 15). The main difference between commercial and non-commercial users usually lies in the requirements for obtaining a licence: commercial users usually pay (more).

It will be possible to set an end date for user accounts, e.g. if a user signs a 30-day evaluation licence.

An interesting topic for discussion is the need for negative access rights (groups that specify "no access" permissions). It could be much easier to use a negative group for e.g. all wave files of the CGN corpus than to use many groups to specify access to all other CGN data (implying access to all but the wave files). This issue is – at the time of writing – open to debate.

Although not necessarily part of the DAM-LR project, we will also discuss how FTP and telnet fit in this access management scheme. Most users will use an IMDI portal to view and/or download data. Other materials, like some corpora, are only accessible via telnet. Any implementation of the access rights proposed in this document should take into account the various types of access. Access rights could then be summarised as this (where S means service):

All CGN data: U:RemcoVanVeenendaal,G:CgnFullLicence RW S:telnet

When all users from an organisation (e.g. the MPI) are granted access to e.g. the CGN, it may be easier to work with IP ranges or domain names in stead of user names:

All CGN data: U:132.229.188.0 – 132.229.188.100 R

or

All CGN data: U:domain="mpi.nl" R

One final remark: following industry best practices, individual users should not receive any access rights. Instead, they should be made member of groups. At the group level, the access rights are set. The sum of the access rights of the groups the user is member of is the individual user's set of rights. In day-to-day use of the system, users will never see the details of their group membership (although there may be some way of requesting this information). They will have access or see a request to sign (and, if required, pay for) a licence.

[15] Current procedure: users print, sign and mail a licence. (If required, they pay a fee.) In return, they receive a user name and password that enables them to either download a LR or view the LR online (the local solution will probably replace the current online catalogue). If a user uses the data for non-licensed purposes, legal action will be taken. The decision to accept a license, grant a certain level of access and/or to sue is not part of a software system (at the INL).

[16] The metadata (IMDI: http://www.mpi.nl/imdi) is publicly available for all (our) LRs.

4 INL LRs

As listing all language resources would result in a lot of redundant information, the access rights settings for only one LR are included here. All other LRs are grouped and given access rights in the following categories:

- Corpora
- Lexica/dictionaries/word lists
- Tools
- Documentation

The Dutch Spoken Corpus (CGN) is one of the most important and larger LRs the INL manages. It is used here as an example of the level of detail on which the proposed access rights system can be applied.

4.1 The Dutch Spoken Corpus

The Spoken Dutch Corpus project was aimed at the construction of a database of contemporary standard Dutch as spoken by adults in The Netherlands and Flanders. The intended size of the corpus was ten million words (about 1,000 hours of speech), two thirds of which would originate from the Netherlands and one third from Flanders. In version 1.0, the results are presented that have emerged from the project. The total number of words available here is nearly 9 million (800 hours of speech). Some 3.3 million words were collected in Flanders, well over 5.6 million in The Netherlands. The corpus comprises a large number of samples of (recorded) spoken text. The entire corpus has been transcribed orthographically, while the transcripts have been linked to the speech files. The orthographic transcription was used as the starting-point for the lemmatization and part-of-speech tagging of the corpus. For a selection of one million words, a (verified) broad phonetic transcription has been produced, while for this part of the corpus also the alignment of the transcripts and the speech files has been verified at the word level. In addition, a selection of one million words has been annotated syntactically. Finally, for a more modest part of the corpus, approximately 250,000 words, a prosodic annotation is available.[17]
The CGN consists of data, documentation, COREX[18] and tools.

4.1.1 CGN: Data

In the table below, the group names and data categories are very detailed. Using (or slightly extending) this scheme, it would be possible to set access rights for the tiniest sets of data. At the time of writing, there are no licences for such specific sets of access rights. Users either have access or they have not[19]. But as we attempt to create a future proof access rights specification system, we thought to include an extreme example.

Group name	Data category	Access right
G:Guest	Metadata	R
G:Guest	Any data	-[20]
G:CgnFullLicence	All data	R
G:CgnAnnot	All annot data	R
G:CgnAnnotXml	All annot/xml data	R
G:CgnBptfon	All annot/xml/bpt-fon data	R
G:CgnBptfonCompa	All annot/xml/bpt-fon/comp-a data	R
G:CgnBptfonCompaNl	All annot/xml/bpt-fon/comp-a/nl data	R
G:CgnBptfonCompaVl	All annot/xml/bpt-fon/comp-a/vl data	R
G:CgnPri	All annot/xml/pri data	R
… (See CGN directories) …	… (See CGN directories) …	R
G:CgnTigCompoVl	All annot/xml/tig/comp-o/vl data	R
G:CgnAnnotText	All annot/text data	R

[17] The English documentation of the CGN is online available at http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm.
[18] The corpus exploitation software developed for the CGN.
[19] Creating tailor-made versions (subsets) of the CGN is also possible (and common practice at the INL). These versions could be added as new LRs, each with their own set of access rights.
[20] A '-' implies "no access".

G:CgnAwd	All annot/text/awd data	R
G:CgnAwdCompa	All annot/text/awd/comp-a data	R
G:CgnAwdCompaNl	All annot/text/awd/comp-a/nl data	R
G:CgnAwdCompaVl	All annot/text/awd/comp-a/vl data	R
G:CgnFon	All annot/text/fon data	R
… (See CGN directories) …	… (See CGN directories) …	R
G:CgnWrdCompoVl	All annot/text/wrd/comp-o/vl data	R
G:CgnLexicon	All lexicon data	R
G:CgnFreqlists	All lexicon/freqlists data	R
G:CgnLexText	All lexicon/text data	R
G:CgnLexXml	All lexicon/xml data	R

Using the table above, user RemcoVanVeenendaal could be made a member of the CgnAnnotText and CgnLexicon groups to have access to the data in annot/text and to the lexicon files. All other data (except the metadata) would be inaccessible.
Granting people access to e.g. all Dutch .pri files is a direct extension of this scheme: users would be member of the groups CgnPriCompaNl through CgnPriCompoNl. Or the users could be made member of a new group AllCgnPri with groups CgnPriCompaNl through CgnPriCompoNl inside (which reflects normal system administration procedures and is more future proof). Predefining all possible groups is impossible; adding groups (and users) must be a feature of any software system implementing the system described in this document.

4.1.2 CGN: Documentation

The documentation of the CGN corpus (protocols, evaluation reports, etc.) may have to be categorised (as evaluation reports may have other access rights than user documentation).

Group name (options)	Data category	Access right
G:Guest	Metadata	R
G:Guest	Any documentation	-
G:CgnProtocols	All protocols within documentation	R
G:CgnReports	All reports within documentation	R
…	…	…

Please note that it is also possible that all documentation should be made publicly available. In that case, all documentation will simply have R for Guests or – slightly more restrictive – an R for CgnFullLicence.

4.1.3 CGN: COREX

COREX is the exploitation (browse and search) software for the CGN. The distribution and versioning of the source code is not an issue for DAM-LR: this will be dealt with outside this platform.
The binaries (executable version) of COREX can only be used with the CGN corpus (locally) available, so distributing COREX separately from the corpus makes no sense. If required, special versions of COREX, working on sub-sets of the CGN, can be developed and distributed, again outside the DAM-LR platform.

4.1.4 CGN: Tools

The tools, developed while creating the CGN, can be made available to the public in the same manner as the CGN documentation: per (set of) tool(s). If all tools are to be made publicly available, set R for Guests (or CgnFullLicence).

Group name (options)	Data category	Access right
G:Guest	Metadata	R
G:Guest	Any tool	-
G:CgnToolsOrig	All tools created during the CGN project	R
G:CgnToolsTST	All tools created after CGN project (by INL or third parties)	R
…	…	R

4.2 Corpora

4.2.1 The 5, 27 and 38 million words corpora

The INL manages three very similar corpora: the 5, 27 and 38 million words corpora
The 5, 27 and 38 million words corpora consist of one or more (ASCII) files with little text structure. Some metadata is available.
Another corpus is the Parole corpus, which is a collection of present-day Dutch texts, containing around 20 million words. The texts were obtained from various publishing houses and other third parties, which implied that their use was to be contractually defined (copyright). Use is permitted for non-commercial research purposes only, and access is restricted to rather small texts fragments, with proper reference of the source.) Some metadata is available.
The NL-Translex project resulted in three spin-off text corpora: Dutch, English and French. At the time of writing, the IPR issues have yet to be sorted out. We do include these corpora in our inventory, but it might be possible that we will not be allowed to grant anyone access.

Group name (options)	Data category	Access right
G:Guest	Metadata	R
G:Guest	Any corpus	-
G:M5Corpus	5 million words corpus data	R
G:M27Corpus	27 million words corpus data	R
G:M38Corpus	38 million words corpus data	R
G:ParoleCorpus	Parole corpus data	R
G:NlTranslexCorpusDutch	The Dutch NL-Translex corpus	R
G:NlTranslexCorpusEnglish	The English NL-Translex corpus	R
G:NlTranslexCorpusFrench	The French NL-Translex corpus	R

4.3 Lexica/dictionaries/word lists

The lexical databases RBN (Dutch reference list), RBBN (Belgian-Dutch reference list), GB95 (word list of the Dutch language, 1995), GB05 (word list of the Dutch language, 2005), ONW (Old-Dutch dictionary), e-Lex (electronic lexicon) and the TST-m-lex (electronic multi-word lexicon) exist as single files (databases). The Parole project also had a (significant) lexicon as a deliverable. Although it might be possible to create finer-grained access rights sets, we only foresee “access or no access” for these files at the moment.
The neologisms list of the Common Dutch Dictionary (ANW) is implemented as a database and will be made available in the local solution as such: a single file with some metadata.
There are several bilingual lexicons resulting from the NL-Translex project: Dutch-English, Dutch-French, French-Dutch and English-Dutch. Each lexicon is an XML file. As with the NL-Translex corpus, IPR issues have yet to be sorted out.
Other files with bilingual data – LRs for translation purposes – are: Arabic–Dutch and vice versa, Danish–Dutch and Indonesian-Dutch. More bilingual files will be available in the future: Dutch-Estonian and vice versa, Dutch-Greek and vice versa, Dutch-Finnish and vice versa and Dutch-Turkish vice versa. These bilingual LRs stem from the ALVV (advisory committee for translation resources).

Group name (options)	Data category	Access right
G:Guest	Metadata	-
GGuest	Any lexicon/dictionary/word list	-
G:Rbn	RBN database	R
G:Rbbn	RBBN database	R
G:Gb95	GB95 database	R
G:Gb05	GB05 database	R
G:Onw	ONW database	R
G:Elex	E-LEX database	R
G:TstMLex	TST-M-LEX	R
G:ParoleLexicon	The Parole lexicon	R
G:NLTransLexDutchEnglish	The NL-Translex Dutch-English lexicon	R
G:NlTransLexDutchFrench	The NL-Translex Dutch-French lexicon	R

G:NlTransLexFrenchDucth	The NL-Translex French-Dutch lexicon	R
G:NlTransLexEnglishDutch	The NL-Translex English-Dutch lexicon	R
G:AlvvArabicDutch	Arabic–Dutch–Arabic data	R
G:AlvvDanishDutch	Danish–Dutch data	R
G:AlvvIndonesianDutch	Indonesian–Dutch data	R
G:AlvvDutchEstonianDutch	Dutch-Estonian-Dutch data	R
G:AlvvDutchGreekDutch	Dutch-Greek-Dutch data	R
G:AlvvDutchFinnishDutch	Dutch-Finnish-Dutch data	R
G:AlvvDutchTurkishDutch	Dutch-Turkish-Dutch data	R
G:AnwNeologisms	Neologisms data of ANW	R

The creation and distribution of "tailor-made" versions (subsets) of these files may become an issue for DAM-LR. See footnote 19 on page 81.

4.4 Tools

Many projects did not only create a LR, but also delivered (exploitation) tools. Some of these tools will/can be included in the local solution. Others are too system-dependant or outdated. Please note that at the time of writing, we only foresee the distribution of binaries via the local solution. The distribution and versioning of source code will be dealt with elsewhere[21].
OMBI (reversing tool for bilingual dictionaries) and documentation could be included as summarised in the table below.
Indexing and retrieval software was created for the 5, 27 and 38 million words corpora. This software was written in Vax Pascal. The retrieval software (with UI) was written using the SGM (screen management) library. It is highly unlikely that users would like to reuse these resources. Also, making the binaries available in the DAM-LR portal seems pointless as they are too system-specific.
The exploitation software for the Parole corpus mainly consists of Perl scripts (for security) on the server side and JavaScript on the client side. The (knowledge of building the) current Parole website could be (re-)used in the DAM-LR project (why invent the wheel again) with minimal effort (since the INL developed the software), but there are better ways of giving users access to the source code than via the local solution.
Finally, although the INL has a licence to use the NL-Translex machine translation system, it is not known when or if the binaries and/or the source code are going to be part of our LR archive.

Group name (options)	Data category	Access right
G:Guest	OMBI and documentation	-
G:Ombi	OMBI and documentation	R

4.5 Documentation

4.5.1 NL-Translex: Documentation

Documentation, like user manuals, of the NL-Translex project and other projects is available. It may be possible that all (user) documentation is made publicly available (R for Guest). Another possibility is that some documentation (evaluation reports) is made available via a separate group (G:ProjectNameReports R).
An example[22]:

Group name (options)	Data category	Access right
G:Guest	All documentation of NL-Translex	-
G:NlTranslexDoc	All documentation of NL-Translex	R

5 Future language resources

In addition to the LRs mentioned earlier, the INL will manage more LRs in the (near) future. Not only as a result of being a central repository for LRs resulting from government-funded projects, but also

[21] E.g. CVS (concurrent versions system) or SVN (subversion: http://subversion.tigris.org/).
[22] Most LRs have some kind of documentation, not just NL-Translex.

because other parties are starting to see the benefits of outsourcing maintenance, management and distribution work to the INL.
Some examples of possible future LRs are:

- Demo versions of LRs. These should be made publicly available (G:Guest R)
- Terminological lexicons of CoTerm (committee for terminology).
- Terminology extractor of CoTerm. This tool (currently under development) will fit nicely in the tools section.
- E-ANS (electronic version of the Dutch grammar rules)
- Regional dictionaries
- Results of STEVIN[23] projects
    - D-Coi: a pilot project for a 500 million words text corpus
    - JASMIN-CGN: extension of the CGN with speech of elderly, children and non-native speakers.
    - COREA: coreference resolution for extracting answers
    - IRME: identification and representation of multi-word expressions
    - AUTONOMATA: grapheme-to-phoneme LRs for proper nouns
- A corpus of 14th century Dutch text
- The Eindhoven corpus (or corpus Uit den Boogaart)
- Spelling web service. The spelling software will be a service available on the INL website. It is unclear is the source code (or data) will be made available via the local solution.

6 Supplementary thoughts

In the past year, the INL has developed a demo version of the CGN corpus (see Future language resources. This demo contains a subset of the data of the entire corpus. Since it is a demo version and publicly available, it might be possible to make this demo guest-readable. This would present visitors of the local solution with the opportunity to preview and/or evaluate the CGN corpus.
All materials without metadata should have metadata created[24] to make browsing and searching the online catalogue of LRs easier.

7 Conclusion

This first step towards a local solution for the INL LRs offers a simple, but practical description of a system of access rights that can be used to implement a tool for managing user access to the INL resources. We have presented examples of how the scheme can be applied to small subsets or entire collections of the LRs and we have given some insight into the many (types of) language resources at the INL.
Armed with the information in this document, we will investigate if we can use available tools to implement a local solution or if we have to create a solution from scratch.

[23] Essential electronic language and speech resources for Dutch: http://taalunieversum.org/taal/technologie/stevin/. At the time of writing, only the first STEVIN call has been issued.
[24] IMDI metadata will be created with the IMDI Editor: http://www.mpi.nl/IMDI/tools

Appendix I

SOAS information architecture and security

SOAS information architecture and security

Data is stored according to OAI model. Structure and filenames of deposits are recorded, then filenames are regularised and used as handles to unique filenames stored in flat directories (one directory per deposit, named with mnemonic of depositor name). Metadata is stored in a custom MySQL database.

ELAR actually makes use of two servers. The main HRELP website (www.hrelp.org) server is externally provided by commercial provider Blackfoot UK. Both the Blackfoot server and LAH run password-protected MySQL databases. The Blackfoot-hosted database and website is concerned with administrative functions of the organisation, whereas LAH's database is dedicated to the storage and management of archival objects and metadata.

Intending users of the archive (i.e. those wishing to access data) must first register with the HRELP website using their email address as a unique identifier. Each subsequent session involves a secure logon. In order to minimise risk, only the hashes of users' passwords are stored. The CAST block cipher, making use of 128-bit symmetric key encryption, is used for sending messages between LAH and the HRELP server when users log in to the archive, as part of the authentication process. No login details are stored on LAH.

To provide appropriate levels of access and management of data, users are assigned to groups (e.g. ELAR Staff, Depositor, User), which are given appropriate levels of access to archive content and functionality. ELAR Staff can accession and manage deposits and metadata. They can add or modify the metadata fields, metadata groupings and the relationships between them, and can modify the rights of other groups and create new groups of users. General users (i.e. those not belonging to any other user group) are able to view archive catalogue metadata, and can obtain copies of deposited materials subject to meeting the requirements for access to those materials.

Appendix J

Lund content formation and access

Lund content formation and access

Over the past year researchers at the Lund Centre have been preparing local linguistic research data for access. These efforts include first and foremost four areas:

- SWEDIA – a phonetic corpus of Swedish dialects, unprecedented in scope and detail
- Swedish and Thai longitudinal child language corpora – approximately half a million running words each plus extensive video linkage
- Archive of Kammu language and culture (sound recordings, drawings, music)
- Online recordings of reading and writing activity (eye tracking, keystroke logging – to be extended with gestures (body tracking))

We aim at getting as much as possible of the above four content resources available on the server – with metadata, authorisation etc – for the workshop/training event in Lund in mid autumn 2006.

Ethical issues – in particular the integrity of subjects in scientific investigations - have recently been under much debate in Sweden. The policies and regulations are contradictory, as illustrated by a case in medicine at Göteborg University, where a researcher was urged to present his research data to the public. In short, it turns out that the rules and regulations of the national research council and associated ethical committees advocate very strict policies including, for example, the subject's right to ask the researcher to destroy the data at any point in time – a policy which would exclude distributing the data on the Internet. On the other hand, Swedish law and the so-called "Offentlighetsprincipen" (Principle of public access) dictates that any citizen has the right to demand to see research data at any point in time.

In response to the situation, the Humanities Laboratory at Lund University has developed strict rules for the handling of research data produced in the lab. The data submitted to DAM-LR partly derive from the Lab and partly from other sources. We would welcome cooperation on how to formulate contracts concerning donors of data which originate from outside the lab.

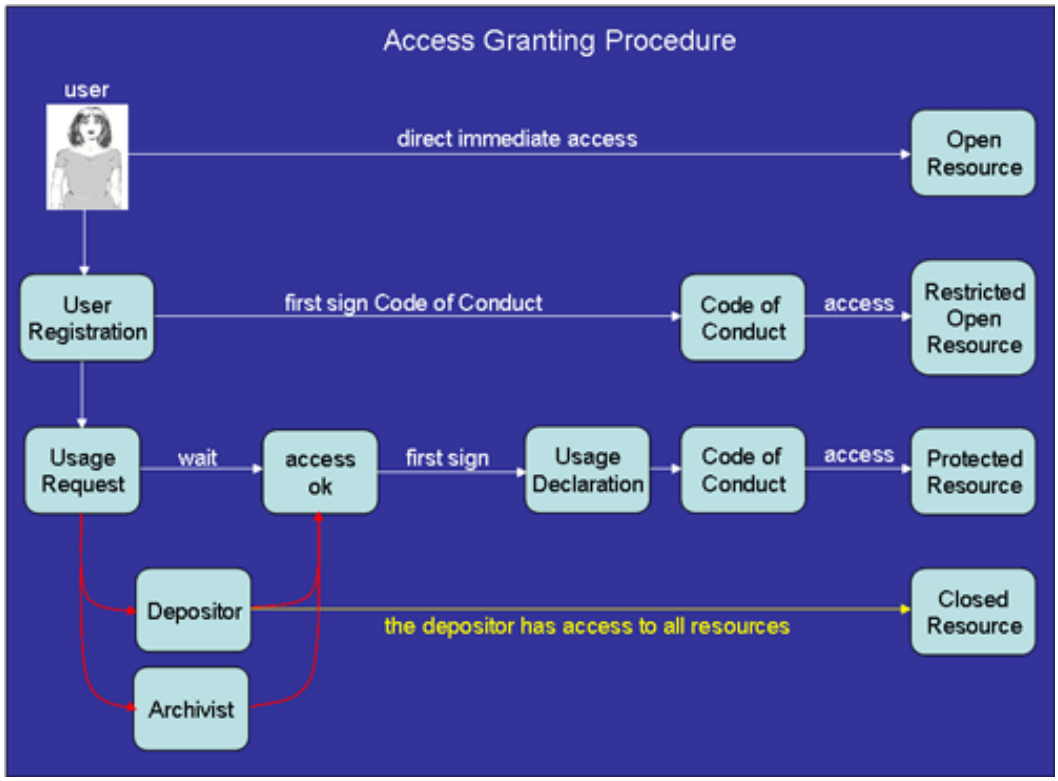
Appendix K

MPI Access Management System

MPI Access Management System

The MPI has established access management principles and built a software system to support it. Here we want to briefly outline these principles and the system.

At the MPI we differentiate between four groups of resources and we are aware of the fact that the resources can change their access characteristics over time. There are resources that are completely open such as pictures etc on the web sites that are meant for the public or for example all metadata descriptions. Of course, general copyright statements are provided to guide the user. Next for most resources we want to ensure that users who want to access the material first sign a " Code of Conduct" before they actually get access. To support this users have to register themselves, i.e., provide a little information about themselves. There are no controls whether the information provided is correct. Next we have resources where the users have to request access by registering themselves and specifying the proposed usage. Either the archivist or depositor will be asked whether access will be granted. If so. they have to sign the usage declaration and the Code of Conduct. Finally, there will be resources that will not be accessible at all.



The MPI archive contains contributions from many different depositors and projects, but for each sub-collection in the archive which can be a singular resource or a large corpus such as the Dutch Spoken Corpus there exists a clear root for which an authorized person is assigned. This person that is normally the depositor or responsible researcher is basically in charge of setting protections, defining CoC etc. Therefore at the MPI we have a number of different CoCs. Dependent on the resource to be accessed the system will check which CoC has to be applied. It is the task of the depositor or the project to define such a CoC. For the DOBES project for example. extensive discussions were held to come to a joint CoC that is valid for all deposits within the DOBES project (see http://www.mpi.nl/DOBES/ethical_legal_aspects/DOBES-coc-v2.pdf). For all those resources or which no CoC was specified by the depositors a "core" CoC was created by the archivist.

MPI's gate keeper software package called LAMUS (Language Archive Management and Upload System) that controls all ingests to the archive has a component called AMS (Access Management System) that allows researchers to define to which category of resources it belongs, to define access groups, to link a CoC, to define authorization managers at all levels within a domain of authority and

to define access rights for a period of time and a certain type of usage. The access rights are very simple: basically authorized persons can assign read rights to one or a collection of resources to other persons. The right to modify resources is not given. Any modification has to be handled via the LAMUS gate keeper which will create new versions and only the authorized person can perform this action.

User authentication is done in two ways: (1) For MPI internal persons the standard ADS user database is copied into an LDAP system specifically designed for archive access management purposes. Since passwords cannot be copied authentication of MPI users finally is still carried out by the central user authentication service. (2) Other users (guests, archive users, etc) can be added to the specific archive LDAP system and they are authenticated by this LDAP system.

This specific LDAP system supports the DAM-LR user attributes and interacts with the Shibboleth identity provider module.