1st Annual 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

Reporting Period: from 01/01/2005 to 31/12/2005

Content

1 Activity Report

1.1 Progress Report

In chapter 1 we want to give an overview about the major activities and achievements of the consortium.

1.1.1 Summary of activities and major achievements

At the Lund days (25-28.1.2006)[1] all items that are relevant for the success of DAM-LR were discussed in great detail. From the presentations it is obvious that the project is on a good time scale:

- Despite the late start of the project all deliverables scheduled for 2005 could be submitted.
- The prototypical system has been fully developed as intended so that we can write the report (D2.2) and integrate the components necessary for the distributed DAM-LR scenario. This means that the MPI is ready, has a completely updated infrastructure and a professional archiving system including all relevant pillars.
- The other partners established archive infrastructures as well and they finished staffing. Dependent on the concrete starting situation and the constraints in their institutions the partners invested great efforts to carry out the archive formation (WP4) and to work on their local archiving system to contain the essential pillars (WP5/6/7).
    - LUND has a full system which is available via the web. However, the server/storage system is a preliminary version. Their main achievements will be done in 2006, but all preparation work has been carried out.
    - SOAS has finished a basic system, including an adequate hardware/storage system, that has all functionality to integrate it in the DAM-LR scenario. They will continue to extend their archiving software in 2006.
    - INL has bought a complete hardware/storage system to house their resources and the intention is to integrate the archiving software in the first months of 2006. Due to its close collaboration with MPI on the archiving software a special installation for the INL was already tested together with the Dutch National Corpus, so that the transformation to the final setup can easily be done.
- With respect to the distributed solution all relevant preparation and specification work has been carried out and the necessary agreements and the architecture were presented and thoroughly discussed at two meetings: (1) the international DELAMAN meeting in Austin (Texas) bringing together a number of well-known archivists and (2) the Lund Days which were seen as a key pillar in the project interaction. Agreements were achieved about the following pillars
    - a complete PKI infrastructure to support certified servers and services
    - an integrated metadata domain based on the IMDI set
    - an integrated system for resolving unique resource identifiers that is based on the Handle System and where every partner is free to define his own postfix syntax
    - specifications for a distributed user management where each partner can chose his own implementation and where Open LDAP is used for the prototypical solution
    - agreements on the user attributes that have to be exchanged within DAM-LR partners and an architecture that allows departments to operate independent of computer centers to fulfill the requirements
    - agreements about integrating access records in the URID database and the formatrequirements
    - an architecture for a distributed authorization system where Shibboleth is chosen to be a core component, where suitable resource manager and access management components have to be developed

[1] The Lund Days were scheduled for December 2005, but had to be postponed to January 2006 due to severe personal matters of one delegation. Therefore, it is mentioned here as part of the work report. All preparations were done in 2005. The subsequent EC meeting accepted the results of the outcome of the Lund Days.

- In addition a discussion was started about the question what the characteristics of the DAM-LR federation will be. This aspect goes beyond the technical agreements. Here the partners agreed that core rules for ethics and legal matters are going to be extracted from what the partners already have to form the basis of the federation.

The partners carried out already some important dissemination work also at international level. It was obvious that DAM-LR is currently the most influential project in the humanities discipline as far as we know with respect to building a concrete research infrastructure. The DAM-LR partners are currently amongst the most advanced institutions with respect to archiving large amounts of data and lending people access to it via web-based techniques and in the humanities domain they are certainly ahead of others world-wide in the application of Grid technologies.

Summarizing, we can state that the consortium is on schedule despite the late start, that all preparation work has been carried out to start the work on implementing the integrated research infrastructure as it is described in the Technical Annex and that already now DAM-LR seems to have a great impact in our domain.

1.1.2 Consortium management activities

Management effort of all contractors

Participant number	**1**	**2**	**3**	**4**	
Participant short name	MPI	LUND	SOAS	INL	**Total**
Person-months	4.80	1.00	0.40	0.60	6.80

Management and general meetings (where all participants are invited)

Date	**Title/Subject of meeting**	**Location**	**Number of attendees**	**Website address**
July 12/13, 2005	DAM-LR Kick-off	MPI Nijmegen	10	www.mpi.nl/dam-lr/
July 12/13, 2005	EC Strategic Meeting	MPI Nijmegen	4	
August 28, 2005	EC Meeting	Tel conference	4	
July 12/13, 2005	WC Planning Meeting	MPI Nijmegen	4	
January 25/28, 2006	Lund Days	U Lund	12	
February 6, 2006	EC Meeting	Tel conference	4	

Milestones and deliverable achievements

Deliverable / Milestone No	**Deliverable / Milestone Name**	**WP / Subtask No**	**Lead Contractor**	**Planned (in months)**	**Achieved (in months)**
M1.1	Present management plan	1	MPI	1	1
M1.2	Create project website	1	MPI	1	1
M1.3	Create annual report	1	MPI	1	1
D1.1	Management Plan	1	MPI	1	1
M2.1	Present first version of access management prototype	2	MPI	7	6
M2.2	Organise project workshop	2	MPI	2	2
D2.1	Prototype Specification	2	MPI	3	2
M3.1	Achieve integrated metadata domain	3	LUND	12	12
D3.1	Metadata Integration Report	3	LUND	1	1
M4.1	Each partner demonstrates an archival system	4	SOAS	10	12

M5.1	Basic decisions about the type of local solution	5	LUND	3	3
M5.2	Document specifications of local solution	5	LUND	8	8
M6.1	Basic decisions about the type of local solution	6	SOAS	3	3
M6.2	Document specifications of local solution	6	SOAS	8	8
M7.1	Basic decisions about the type of local solution	7	INL	3	3
M7.2	Document specifications of local solution	7	INL	8	8
D8.1	Definition Report T6	8	INL	2	2
D8.1	Definition Report T12	8	INL	8	8

1.1.3 Project activities (Design study, Construction)

The following table gives an overview about the total person-months per project activity for each contractor:

Participant number	**1**	**2**	**3**	**4**	
Participant short name	MPI	LUND	SOAS	INL	**Total**
Management (WP1)	4.8	1	0.4	0.6	**6.8**
Local Prototype (WP2)	35	0	0	(see WP7)	**35**
Metadata Integration (WP3)	1	3	0	7	**11**
Archive Formation (WP4)	31	11	16.8	75	**133.8**
Local Lund (WP5)	0	5	0	0	**5**
Local SOAS (WP6)	0	0	4.2	0	**4.2**
Local INL (WP7)	0	0	0	25	**25**
Definitions (WP8)	2.32	0	0.8	0.1	**3.22**
Distributed Solution (WP9)	3.14	0	0	0.5	**3.64**
Dissemination (WP12)	1.65	3	1	0.4	**5.89**
Total (person months)	**78.91**	**23**	**23.2**	**108.4**	**233.51**

It can clearly be seen that the large investments were made in WP2 by MPI to setup a prototypical solution and by the partners in WP4/5/6/7. Dependent on the local situation and the goals that are going to be achieved by the partners differences with respect to the efforts occurred, of course. For a detailed list of activities see appendix B.

1.2 List of deliverables

During the reporting period the following deliverables are achieved:

Task Nr.	**Deliverable**	**Name**	**WP Nr.**	**Delivered by Contractor(s)**	**Planned (in months)**	**Achieved (in months)**
1	D1.1	Management Plan	1	MPI	1	1
2	D2.1	Prototype Specification	2	MPI	3	2
3	D3.1	Metadata Integration Report	3	LUND	1	1
8	D8.1	Definition Report T6	8	INL	2	2
8	D8.1	Definition Report T12	8	INL	8	8

This list can prove as well that the activities of the DAM-LR project are in time.

1.3 Use and dissemination of knowledge

Already in its beginning year where the setup of the local archives was in the focus a number of occasions were used to present the DAM-LR ideas and get feedback. The following list gives an overview about these events.

Date(s)	Who	Location	Number of attendees	Title, Website address
March 15	MPI	Tilburg	15	LAMUS introduction Dutch Bilingual DB meeting
April 14	MPI	Berlin	>100	Open Forum for Metadata Registries
April 29	MPI	MPI Nijmegen	30	Studiedag University Nijmegen
May 23-27	MPI	MPI Nijmegen	30	DOBES Training course www.mpi.nl/DOBES/training/training.html
May 23,24	MPI	MPI Nijmegen	70	DOBES Conference www.mpi.nl/DOBES/workshop/workshop.html
July 2	SOAS	Harvard, MA	75	EMELD workshop on linguistic ontologies and data categories for language resources http://emeld.org/workshop/2005/
July 5	LUND	Poitiers	15	DAM-LR concept introduced to network of French researchers on writing
August 21-23	MPI	Warshawa	50	ISO TC37/SC4 Meeting
August 23	LUND	Lund	20	First international research workshop ”Brain Mind Behaviour”; DAM-LR as a vehicle of facilitation www.sollu.se/bmb
September 7-8	LUND	Lund	25	International research workshop concerning reading and writing in real time incl DAM-LR as a vehicle of facilitation
October 1	LUND	Lund	30	Presenting Lund Lab including DAM-LR to delegation of Norwegian researchers
October 10-14	MPI	MPI Nijmegen	29	DOBES Training Course
October 11	LUND	Lund	20	Second international research workshop ”Brain Mind Behaviour”; DAM-LR as a vehicle of facilitation www.sol.lu.se/bmb
Oct 21	SOAS	London	12	ELAP/ELAR Research Seminar http://www.hrelp.org/events/seminars/ELAP-ELAR/index.html
October 22-24	LUND	Stanford		Meeting with WGLN incl DAM-LR as a vehicle for research cooperation between Stanford and universities in Sweden
November 9	MPI	University Hannover	15	Germanistik Seminar
November 18	MPI	KNAW Amsterdam	15	Challenges Workshop
November 21/22	MPI	University of Texas, Austin	20	DELAMAN Meeting www.ailla.utexas.org/site/delaman.html
December 12/13	MPI	Berlin	>100	International Standards Conference

1.4 Activities per Partner

Here we briefly give an overview about the partners main activities. For more details we refer to the activities table with all details specified by the partners in appendix A.

1.4.1 MPI

WP1
The MPI carried out all activities that were necessary to start the project, to allow for a smooth interaction between the partners and keep track of all activities. Most important were the agreement on a management plan, the deep interaction with all partners at the kick-off meeting which was used to come to first agreements and the Lund IT days which were used to talk about the basis of our federation and the complete architecture of the DAM-LR infrastructure. The Lund IT days took place in January 06, i.e. they formally do not fall into this reporting period. But all preparations were done in 2005 and due to the late start of the project the small delay is acceptable. In particular after this meeting with some members of the executive committee, most of the Working meeting and some specialists participating we can say that all agreements that are necessary have been done.
Further, we took care that the members of the Executive Board had the chance to comment on the outcome of the two important meetings mentioned. This was done via telephone conferences – one at 25.8.2005 and the other at 6.2.2006. Both meetings yielded full support for the agreements defined in the work meetings mentioned above.

WP2/WP4
The focus of the work was on developing a prototypical solution that can be used as a reference system within DAM-LR. This work included the own contribution of the MPI to the DAM-LR scenario. A prototype specification was produced as deliverable D2.1 that covers all relevant aspects. In the deliverable D2.2 we will give a report that we finished all important work in this respect. A complete archive system was set up that contains now more than 150.000 resource objects that require more than 15 Terabyte of storage capacity. The media archive, i.e. the fully described part of it is completely accessible via the web and in particular the metadata descriptions created in accordance to the IMDI standard are openly accessible. A number of professional tools support browsing, searching, downloading etc.
Another major tool was developed which is the basis for the DAM-LR work: LAMUS is a content management system with special functionality for language resources. We cannot describe all its functionality here, but it is certainly one of the most advanced tools now in the linguistic domain. It comes along with the AMS component which allows archivists and depositors to define access policies and rights. LAMUS was designed and developed in collaboration with INL.
Further, all steps were made to include Unique Resource Identifiers (URIDs), to step over to LDAP as a more advanced user management and authentication system and to integrate Shibboleth all being essential for the DAM-LR scenario.
Finally, the MPI invested much money to create an appropriate storage and networking architecture[2]. In total 444.000 € were spent to buy a powerful router to support media streams to the outside world, new powerful and redundant network switches to guarantee a local network with enough capacity and high availability, a multi-layer storage system with two new SUN servers, a fast RAID with about 4 TB, a slow RAID with about 15 TB to store the online media data and new SAN switches, and a number of multiprocessor LINUX servers to house databases, indexes to support fast searching, web-services and normal Internet exchange.
Summarizing, we can speak about a fully operational and professional archiving setup with a prototypical system as described in the Technical Annex, i.e., the development of the prototypical system and the archive formation at the MPI can be seen as finished (of course, we will have to carry out continuous upgrades). Given appropriate access rights all organized archive content can be accessed via the web.

WP3/WP8
Here we collaborated with Lund university to improve their metadata descriptions and to bring them into the current IMDI version.
In WP8 we collaborated with INL to produce the definition reports (D8.1/T6/T12).

WP9
With respect to WP9 we are in a very satisfying state already, since the MPI spent more time than planned to carefully specify the requirements for a distributed scenario, study the possible components, test most of them and discuss them with the partners and at other occasions such as for

[2] Due to budget reasons some investments within the DAM-LR framework had to be done in 2004, but the installation and integration of the equipment was done in 2005.

example the DELAMAN meeting (an international network of language resource archives, www.delaman.org). The result was an extensive report about all aspects that will have to be tackled by DAM-LR to come to final operational research infrastructure. This report is included in appendix B. It was the basis for the discussions at the Lund days where we achieved concrete agreements on almost all points. This will then be part of the official specification report which is due in March 2006. So all necessary work has been done in this respect as well, so that the implementation, test and integration work can be started as planned.

WP12
As described in chapter 1.3 the MPI was very active in disseminating the ideas of DAM-LR and to get feedback in particular at an international level. In addition, we started a discussion procedure in the framework of the ESFRI roadmap plans to address members of all European member states to build up a EU-wide research infrastructure. It is obvious that DAM-LR results can and should play an important role in designing such an infrastructure – it would mean to transform DAM-LR approaches to a much larger domain and demonstrate their scalability.

Investment Summary
In total MPI invested in DAM-LR the following:

- about 444.000 € to update and optimize its local compute, storage and network infrastructure;
- (more than) 71 person months of its own staff to extend and maintain the archive, to optimize its management procedures by improving its coherence and consistence, to design, develop and test the prototypical system, to look into the components needed for DAM-LR, and to do the other DAM-LR related activities;
- 8.11 person months for DAM-LR related tasks funded by the EU of which 3.8 person months were spent on management activities.

1.4.2 Lund

WP1
Lund also carried out some management activities in the management area. Basically these were to interact at high university level and beyond to get the funds for the archive formation part, to form a local archiving group and to prepare the Lund DAM-LR meeting that took place in January 2006.

WP3
Quite some work was done to improve the existing Lund archive information basis, to clean the metadata descriptions and to transform them to the latest IMDI version. Recently, all metadata was integrated into the newly setup first archive solution.

WP4/5
Most work was spent of course to setup a local Lund archive in all respects. (1) A completely new building was provided and designed so that it can house the new Lund language resource archive as well. In parallel, the whole department was restructured so that it not only houses anymore the linguistic department, but other departments of the humanities as well with the goal to create a highly communicative atmosphere where various disciplines can take profit from the built-in infrastructure such as the emerging language resource archive. This restructuring was basically finished in 2005.
Part of this restructuring was the establishment of an “archiving” framework for the faculty. The plans were almost completely finished in 2005, i.e. the formation of an archiving working group took place, a phase plan was developed for building up a self-standing computer, storage and network solution, training courses were held about metadata and archive construction and the realization of the first small solution was planned. In collaboration with the MPI a first language resource archive was finally set up in January and is operating so that a number of well-described resources are accessible via the web now. The solution is based on the prototypical solution developed at the MPI and therefore contains all necessary pillars.
Plans in the faculty were made such that other resources will be integrated into the archive as well. It was estimated that within the coming five years a capacity of about 100 TB would be required in particular due to the digitization of media streams and due to the many time series data recorded in the various labs. This allowed Lund university to make plans for buying a full fledged multi-layer storage solution in 2006 and to request the necessary amount of funding (as was specified in the DAM-LR proposal). The budget request was finally granted so that achievements can be made in

2006. Appropriate lab solutions were planned as well to support the flexible integration of lab data into the archive.

WP12
Lund also participated in dissemination activities. Several events were organized in Lund and DAM-LR ideas were also raised at international meetings.

Investment Summary
In Lund all activities in 2005 were taken care of by Lund staff members, i.e. Lund will start spending European funding in 2006. The total investments from Lund were as follows:

- about 75.000 € for the preparation of the local system which will be bought and installed in 2006
- about 23 person months of its own staff for all activities mentioned above

1.4.3 SOAS

WP1
SOAS took part in the preparation of the various meetings at DAM-LR and internal level.

WP4/6
As for all partners the focus for SOAS lay in the setup of the local archive in all respects. The ELAR division was established at SOAS to take care of all archiving matters that occur due to the HRELP programme. An organizational framework was set up and the stuffing of the digitization and archiving division was carried out. Further, agreements were done with the Oxford Text Archive about a second copy for long-term persistency purposes.
A complete architectural design was done and presented first at the DELAMAN meeting in Austin. A basic infrastructure was set up containing digitization equipment, powerful servers, storage and networking components. Some building work had to be carried out to house all components. To store some off-line material a safe was installed as well. Much work went into establishing proper workflow schemes that could guide the users from the recording to the depositing phase.
The archive management software was built to a certain extent so that first uploads could be made. Deposit forms were created so that depositors can specify their upload wishes. A metadata set was designed and implemented that is used to describe the resources. A solution was designed that allows to create different types of metadata records, in particular IMDI records that are used within DAM-LR. In addition, the Handle system was installed to include URIDs in the archiving scheme at SOAS from the beginning. A local syntax scheme for the handle postfixes was designed.
The user administration and authentification is done with the help of a separate database that is integral part of the SOAS archiving software.

In general, it is obvious that in 2006 the SOAS archive will be ready to become an integral part of the DAM-LR scenario as is intended.

WP8
SOAS contributed to the establishment of the definition reports.

WP9
SOAS took the lead in DAM-LR to work on the question what a DAM-LR federation will be beyond the agreements enforced by technical considerations (see appendix C). A talk with important points was given at the DELAMAN meeting in Austin and the points were put on the agenda of the Lund meeting. Further, SOAS contributed to the discussions about the architecture of the distributed solution.

WP12
SOAS used two international meeting and a training course to spread information about the DAM-LR project and its goals.

Investment Summary
SOAS made the following investments in 2005 in relation with the DAM-LR goals:

- about 301.000 € were spent to setup the local compute, storage and network infrastructure and to carry out the necessary building works etc

- (more than) 13.2 person months of its own staff were spent to setup and manage the new archive infrastructure and for the other DAM-LR related activities

1.4.4 INL

WP1

INL carried out some management activities in the realm of the DAM-LR project. In particular INL took care of the Wiki for DAM-LR which is used to store and exchange relevant documents and to comment on them. Also an internal DAM-LR working group was formed containing specialized persons from the INL and the TST Center staff.
It should be mentioned that Remco van Veenendaal was officially nominated by INL as alternative member of the EC in case that Jeannine Beeken cannot participate.

WP2

INL was heavily involved in the specification and realization of the LAMUS part of the prototypical system that was developed at the MPI. Much effort was invested to design and implement different APIs and classes that are now used within LAMUS. Tests were carried out to determine whether LAMUS can be used within the TST Center and INL. The details were already mentioned in chapter 1.4.1.

WP3

In collaboration with the MPI the metadata descriptions of the large national Dutch Spoken Corpus were transformed to the lastest IMDI versions. They are all ready to be integrated into a local domain. The whole corpus is now available in a well-described manner in open and harvestable formats.

WP4/WP7

INL has invested an enormous effort to set up the TST Center – a Dutch-Flemish center that is going to store a wide variety of textual and speech resources, in particular in Flemish and Dutch, and offer services to interested parties. The TST Center was set up in 2004, i.e., accommodation and staffing aspects were widely solved. A technology team that was formed in close collaboration with the technical team of the INL had to design requirements for language resource archiving. With respect to setting up the local archive the following activities were carried out:

- writing a full specification of the required server, storage and network requirements
- setup and maintenance of the infrastructure, which is now fully operational
- web-sites were specified and implemented that allow access to the resources
- a preliminary user management system was implemented as well to start offering services
- the offering of services started
- specification of the requirements for access management given the specific circumstances with respect to property rights and procedures
- conversion, in collaboration with MPI, of the whole Dutch Spoken Corpus, to fit with the new LAMUS archive management software and its integration into the archive
- integration of the converted Dutch Spoken Corpus into the archive

It can be stated that all preparations have been carried out to establish a DAM-LR compliant archive setup at TST center in the first months of 2006 and to be able to offer a fully operational services to the customers.

WP8

Two versions of the definition reports (D1) were produced in collaboration with the partners. The definitions were based on the agreements achieved at the kick-off meeting and by email and telephone interaction. A new version will be created that covers the agreements achieved at the Lund days.

WP9

Some evaluation work was carried out to assure the quality of the definitions and architectural plans with respect to the distributed solution. To be mentioned are: access policies and rights, URID resolving architecture and syntax and metadata definition and harvesting concepts. In focus are always the security aspects.

WP12

Some dissemination activities were carried out, essentially the preparation of an LREC workshop to happen in 2006.

Investment Summary
INL is working under an FCF cost model in contrast to the other partners. The following investments were done by INL in 2005:

- about 48.000 € was spent to setup the local compute, storage and network infrastructure
- (more than) 102 person months of its own staff were spent to setup and manage the new archive infrastructure, to co-develop the archiving software and to prepare its installation
- 5.8 person months for DAM-LR related tasks funded by the EC

2 Management Report

In this chapter we want to give an overview about efforts that were taken by every partner that were contributed by themselves to the project.

2.1 Total additional costs (non-personnel)

Who	Costs (k€)
MPI	444
LUND	75
SOAS	301
INL	48
Total	**868**

2.1.1 MPI additional costs

Description	WP	Costs (k€)
Powerful router to support media streams	4	10
New powerful and redundant network switches	4	105
Storage systems (RAID, Servers, Switches)	4	249
Various servers (Databases, search, web, etc)	4	64
Various small investments (digitization etc)	4	16
Total		**444**

2.1.2 Lund additional costs

Description	WP	Costs (k€)
Preparations for major investments 2006 (storage and data providing systems)	4	75
Total		**75**

2.1.3 SOAS additional costs

Description	WP	Costs (k€)
Storage area network (large disk arrays)	4	42
Robotic tape library	4	40
Server	4	3
Data safe	4	43
Office furniture	4	14
Dobbin Audio Workflow system	4	19
Building works	4	112
OTA agreement	4	28
Total		**301**

2.1.4 INL additional costs

Description	WP	Costs (k€)
Servers and storage	4	35
Laptops	4	10
Software	4	3
Total		**48**

2.2 Total additional costs (Personnel)

In this chapter we want to summarize the personnel investments that each partner brought in. In particular, these were part of the archive formation process as described in WP4. Often this cannot clearly distinguished from the setup of the local archive as described in WP5/6/7. As can be seen from the activities table (see appendix A) each partner also invested own personnel to work on issues mentioned in other work packages. To give a rough indication of the costs involved we estimate the average costs of a qualified person of 60.000 € per year.

WP No	WP Title	MPI (pm)	LUND (pm)	SOAS (pm)	INL (pm)	Total (pm)
WP1	Management	1.0	1.0	0.4	0.1	**2.5**
WP2	Local Prototype	35.0			see WP7	**35.0+**
WP3	Metadata Integration	1.0	3.0		6.0	**10.0**
WP4	Archive Formation	31.0	11.0	15.8	72.0	**129.8**
WP5	Local Lund		5.0			**5.0**
WP6	Local Soas			3.0		**3.0**
WP7	Local INL				24.0	**24.0**
WP8	Definitions	0.2		0.8		**1.0**
WP9	Distributed Solution	2.0		1.2	0.5	**3.7**
WP10	Adapt and Integrate					
WP11	Test					
WP12	Dissemination	0.6	3.0	1.0	0.2	**4.8**
Total pm		**70.8**	**23**	**23.2**	**102.8**	**218.8**
Total costs (k€)		**425**	**138**	**139**	**617**	**1319**

2.3 Total additional costs (personnel + non-personnel)

In this table we give a total overview about the own investments of the partners and compare this with the own contributions that were estimated in the proposal and that are included in the TA.

Who	Non-personnel Costs (k€)	personnel Costs (k€)	Total Costs (k€)	Estimated Costs in TA (k€)
MPI	444	425	869	1061
LUND	75	138	213	790
SOAS	301	139	439	740
INL	48	617	665	980
Total	**868**	**1288**	**2156**	**3571**

After the first year we can see that two institutions (MPI, INL) are already very close to what they have estimated as own contributions. MPI and INL are already fairly far in setting up their local archives. Lund will invest in a complete server, storage and network architecture in 2006 and estimates costs of about 400 k€ for this action. SOAS will still invest quite some money to improve their local archiving software, so here the big investments will be on personnel costs. In total we guess that after the total three years period of the DAM-LR each partner will have invested more than the estimated costs. So, our estimated numbers indeed to be an excellent base line.

3 Report on the distribution of the community financial contribution

The community financial contribution (pre-financing) was distributed to the contractors according to the payment modalities (article 8) of the DAM-LR contract in time. The table gives an overview how the received financial contribution from the EC of about 92.000 € was distributed amongst the partners (see last column).

	2005	**2006**	**first six months of P2**	**first 18 months**	**pre-financing**
	P1	P2	P2/2	P1 + P2/2	0.8 * (P1 + P2/2)
MPI	24	51	25.5	49.5	39.6
LUND	4	37	18.5	22.5	18
SOAS	2	39	19.5	21.5	17.2
INL	2	39	19.5	21.5	17.2
Total	**32**	**166**	**83**	**115**	**92**

Appendix A : Project Activities

The table below gives an overview of all the project activities per work package

WP	Who	Activity
1	INL	Setting up the DAM-LR Wiki
1	INL	Formed internal Working Committee
1	INL	Taking part of DAM-LR kick-off meeting
1	LUND	Formed Lund work group
1	LUND	Participation in DAM-LR meetings
1	LUND	Building a local group for organizing training events
1	LUND	Contributing to D1.1 Annual Report 2005
1	MPI	Project administration and coordination
1	MPI	D1.1 Management plan
1	MPI	Project website set-up and maintenance
1	MPI	Formed Executive Committee
1	MPI	Formed Working Committee
1	MPI	Set up email list
1	MPI	Preparation for DAM-LR Kick-off meeting
1	MPI	Organize DAM-LR Kick-off meeting
1	MPI	Preparation for Lund workshop and training (to come in January 2006)
1	MPI	Finish D1.1 Annual Report 2005
1	SOAS	Project WC/EC meetings and preparation
2	MPI	Design of Local Prototype
2	MPI	Implementation of Local Prototype
2	MPI	Optimize “Local” Metadata Infrastructure
2	MPI	Creation and optimization of “Local” Access Management Infrastructure
2	MPI	Testing of Local Prototype
2	MPI	Build and Extend “local” Archive
2	MPI	Finish D2.1 Prototype Specification
3	SOAS	Define and implement a metadata scheme
3	SOAS	Design exchange wrapper of metadata scheme with MPI
3	INL	Discuss extended metadata requirements with MPI
3	INL	Integrate CGN metadata
3	LUND	Optimize “Local” Metadata Infrastructure
3	LUND	Analyse and prepare to integrate Lund metadata
3	MPI	Analyse and prepare to integrate Lund metadata
4	INL	Buy and set-up storage and server systems
4	INL	System management task
4	INL	Archive management task
4	INL	Setup the local archive infrastructure
4	LUND	Creation and optimization of “Local” Access Management Infrastructure
4	LUND	Setup the whole lab environment for the local archive
4	LUND	Setup the “local” Archive infrastructure
4	LUND	Preparations for major investments 2006
4	LUND	Lund XML-project for online recordings of speech, writing, and visual and tactile reading
4	MPI	Buy and setup storage and server systems
4	MPI	Upgrade network
4	MPI	Copy data to new storage and servers
4	MPI	System management task
4	MPI	Archive management task
4	SOAS	Storage architecture design
4	SOAS	Storage systems implementation
4	SOAS	Tape library scheduling & testing
4	SOAS	Safe research and order
4	SOAS	Create equipment planner, archive section, specifying and ordering
4	SOAS	Research and broker Oxford Text Archive replication agreement

4	SOAS	System and network management tasks
4	SOAS	Set up Handle system
4	SOAS	Research, order, install Storage Area Network, Tape Library, Server
4	SOAS	Setup the local archive infrastructure
5	LUND	Install local archiving software
5	LUND	Testing of Local archiving software
5	LUND	Add available resources to the archive and make them web-accessible
6	SOAS	Archive architecture review
6	SOAS	Prepare deposit form
6	SOAS	Archive website development
6	SOAS	Prepare and install signage and artwork
6	SOAS	Design and oversee building works
6	SOAS	Research and formulate collection and accession policies
6	SOAS	Design pre-ingestion workflow
6	SOAS	Design and install the local archive software
6	SOAS	Carry out first archive uploads
7	INL	Design of local prototype
7	INL	Implementation of local prototype
7	INL	Preparation of local metadata infrastructure (IMDI-portal)
7	INL	Preparation of local access management infrastructure
7	INL	Install and test the local archiving software on a preliminary server
7	INL	Developing the alpha version of the local archive
8	INL	Contribute to and finish definition report 8.1 in cooperation with MPI
8	MPI	Assist INL with definition report D8.1
8	SOAS	Discuss and document local architecture definitions
9	INL	Inventory and proposal regarding access policies and rights
9	INL	Evaluation of URIDs
9	INL	Discuss metadata aspects and harvesting concepts
9	MPI	Analyse and report about A&A solutions
9	MPI	Analyse and set-up Handle System
9	MPI	Analyse and report about URIDs
9	MPI	Define URID architecture and specification
9	MPI	Analyse EUGrid PMA solution
9	MPI	Get certificates and set up a PKI system
9	MPI	Analyse Shibboleth and exchange information with experts
9	MPI	Analyse LDAP and exchange information with experts
9	MPI	Design a complete A&A architecture and present it
9	MPI	Discuss metadata aspects and harvesting concepts
9	MPI	Finish the A&A design report
9	SOAS	Evaluate current “federation” discussions and make proposals for a DAM-LR federation
12	INL	Preparation for workshops, panel discussions and papers at the LREC-conference (May 2006)
12	LUND	(Co)organizing meetings and workshops where DAM-LR was presented (see table below)
12	LUND	Preparation for Lund workshop and training course (to come in January 2006)
12	MPI	Preparation for workshops and conferences
12	MPI	Preparation for DELAMAN workshop
12	MPI	Preparation for DOBES conference and training
12	MPI	Preparation for LREC workshop and conference (to come in May 2006)
12	SOAS	Preparation for EMELD conference
12	SOAS	Preparation for ELDP training
12	SOAS	Prepare papers for Delaman III

Appendix B : Report for a distributed setup

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: 13.1.2006

Content

4 1. Introduction

This report is meant to summarize all basics that have to do the with 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 and not yet been decided have to be decided 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 to be 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 accepted. The broadest points in the discussions 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 a larger scope. It is obvious that this also has to be subject of a discussion. 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 will distribute his note on “federations” as soon as possible.

5 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.

6 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 given 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 a its handles. This requires, however, that handle requests crossing the local boundaries have to be resolved by the global handle resolving service. Caching could be used to increase performance.
- Every partner has full control about his Handle database, i.e. no one else will get the permission to change entries except via clearly defined services in the case of modifications of paths for copied data.

[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.

- Every partner therefore has to install and maintain the Handle System on a server and has to take care that its database will be maintained properly.
- For redundancy reasons the MPI will host mirrors for all partner services, i.e. in case of server problems the URIDs could still be resolved.
- There is a recommendation to not use semantics within the postfixes, but in fact every partner is free in his decisions.

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

7 4. Authentification

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.
- It is possible partner institutions/departments do not control their user administration, i.e. they have to start a discussion process of how to best create a joint domain.
- It may be necessary (see below) to have the possibility of one integrated search domain, i.e. it should be possible to propagate some open attributes of users to a trusted higher node such as it is possible with LDAP.

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

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

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

[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.

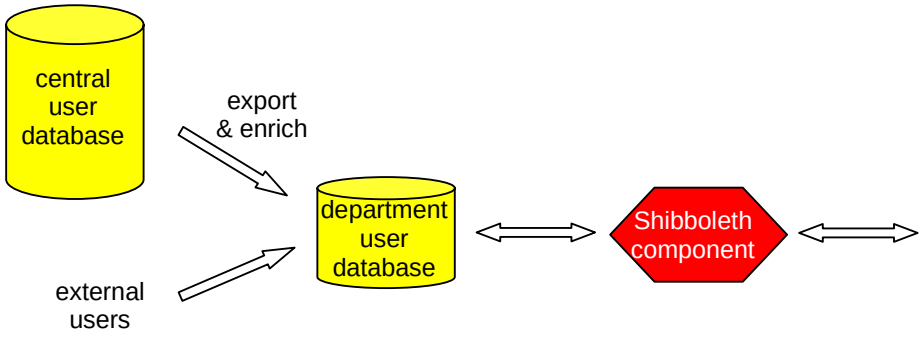
These attributes seem to be the minimal set and are largely overlapping with the specifications from EDU-Person and RFC 2798[6], others such as a host introducing an external guest as a collaborator or so may be necessary but are not yet fully identified. The MPI certainly will store internally aspects such as department, start of employment, end of employment and behavior flag. In DAM-LR we have to decide whether we will speak about accounts that are valid for a limited period of time only and whether this limitation is associated with specific requests and/or with the account itself. Both seems to be appropriate. The information about the duration of a request certainly would have to be stored together with the other request information.

The behavior flag is relevant for the MPI to indicate persons who severely misbehaved. If the flag is set all access will be ignored. We have to have such possibility to memorize such form of misbehavior. Of course, we cannot prevent completely that the same person will register again under another name. However, when we would apply the host concept it would become difficult to sail under other names. This issue is tricky and has to be discussed.

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

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

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



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

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

8 5. Authorization

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

8.1 5.1 General Aspects

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

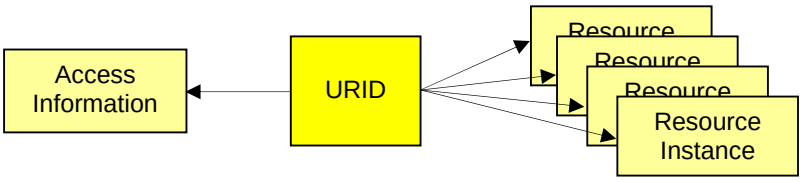
- single identity — to be achieved by distributed authentication and accepting attributes
- single sign-on — once the user is identified he/she has transparent access to all resources he/she is permitted to access in all archives
- one basket idea — the user must be able to see his/her set of accessible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

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

One of the basic agreements in DAM-LR is that access handling is done by the originating institution. Since for each resource independent of the number of instances there will be only one URID[7], it seems to be a direct conclusion that

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

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



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

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

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

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

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

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

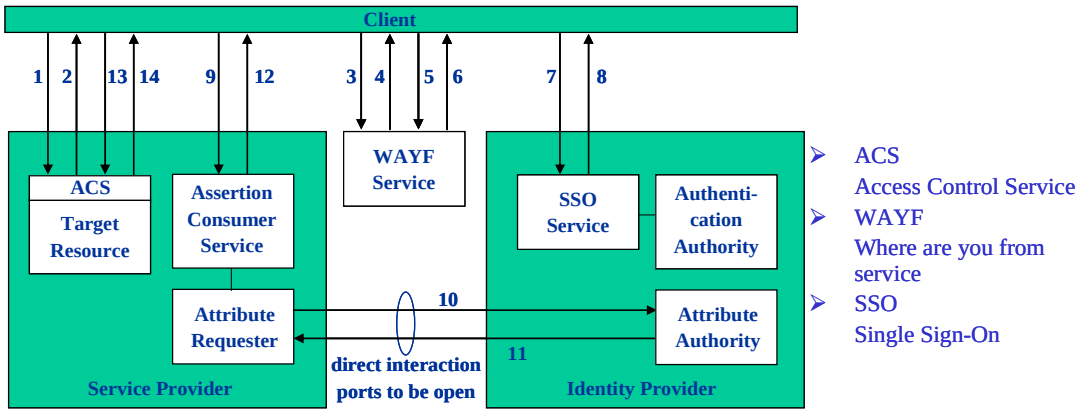
8.2 5.2 Shibboleth

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

Let us therefore first introduce Shibboleth briefly (for details we refer to the Shibboleth documents). The following figure indicates the different Shibboleth components (as described in older documents).

[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.

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



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.

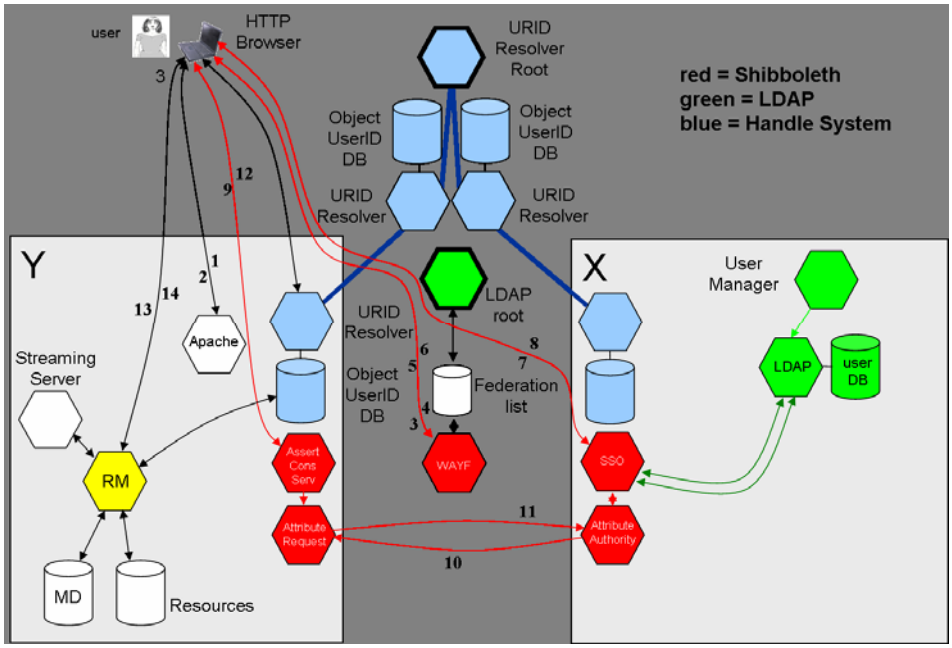
It should be discussed in detail what the best solution will be for us.

8.3 5.3 Typical Access Scenario

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

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

the Resource Manager (13). The resource manager will check whether the rights are ok by interacting with the Object-User Database and finally deliver the requested data.



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

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

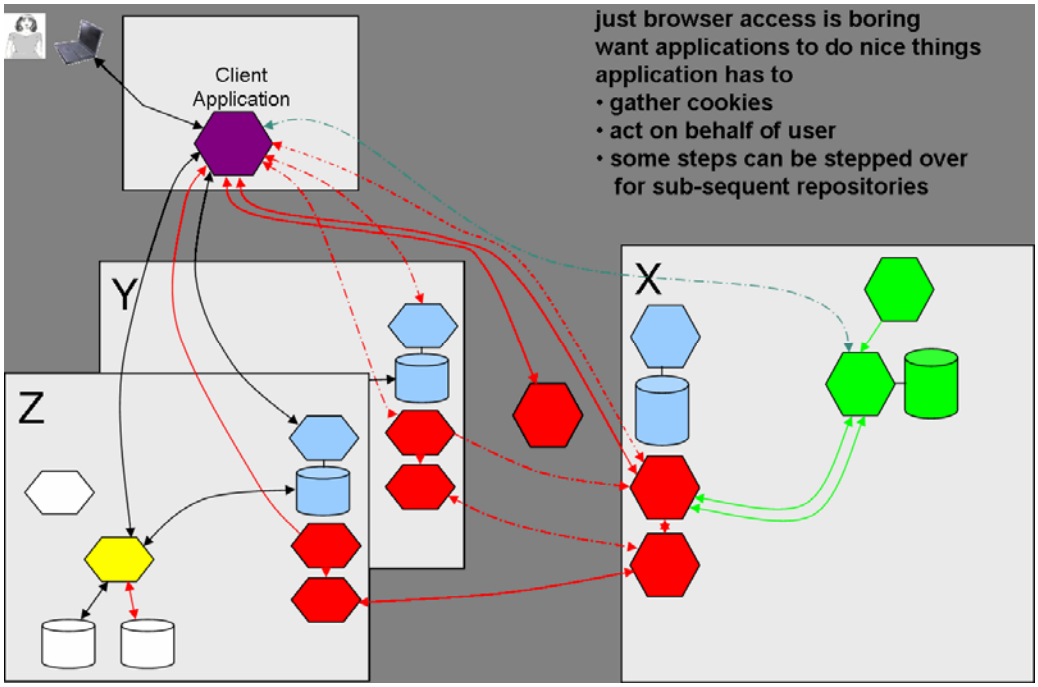
We should look for existing solutions rather than do all the programming ourselves.

8.4 5.4 Application Access

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

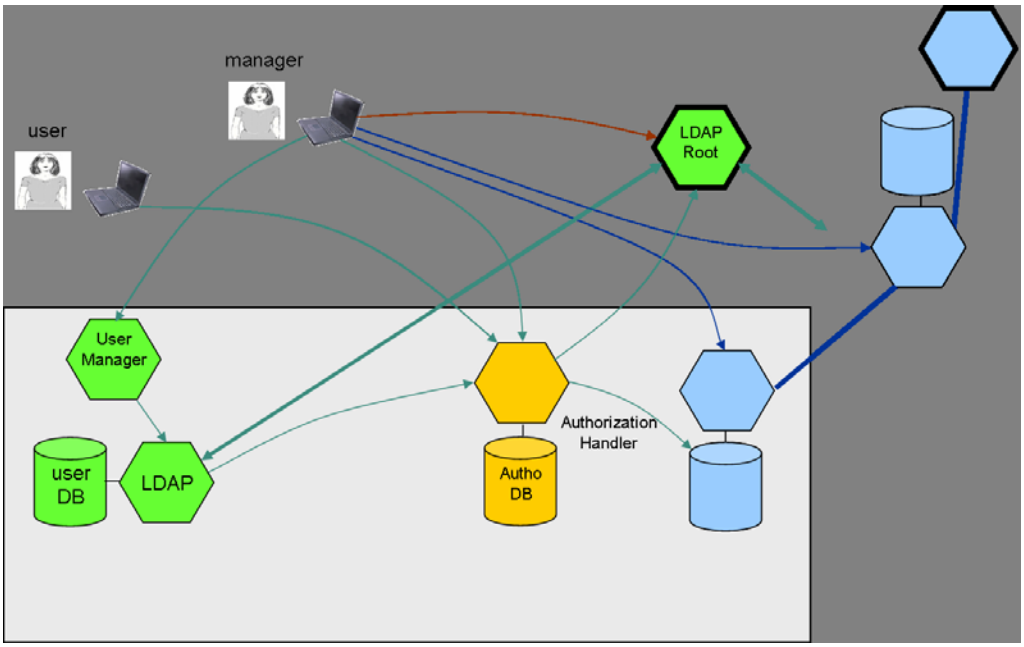
The application has now to act on behalf of the user if different sites are involved. The user name and password will be requested by every site that has to be contacted, but the user should type it in just once. To access several resources from one site the application needs to store the appropriate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

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



8.5 5.5 Management Scenario

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



New User

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

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

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

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

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

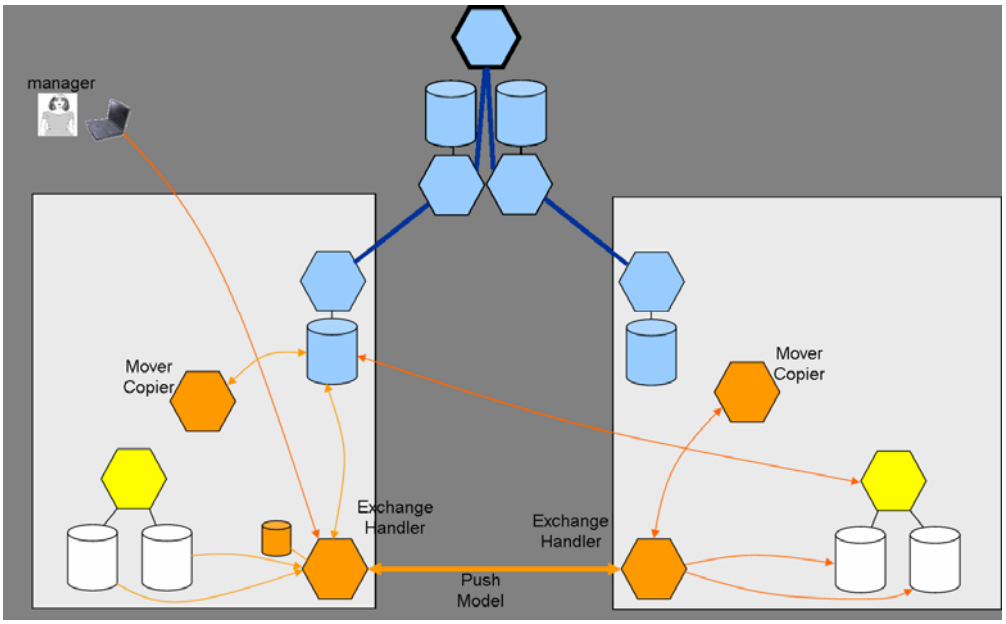
The Authorization DB also contains per sub-corpus specifications about processes such as whether the depositor has to be asked first, whether the person has to sign a declaration etc. When the user has fulfilled all required steps the general command will be transferred into corresponding formal URID access records[8] 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.

8.6 5.6 Data Moving Scenario

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

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



9 6. Summary

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

9.1 6.1 Software Components and Certificates

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

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

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

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

- a Resource Manager as described above
- a lookup routine that extracts from the LDAP root which partners are federation members[10]
- the MPI intends to extend its ANNEX and LEXUS applications to be used as test beds for the multi-resource scenario
- an Authorization Handler that interacts with other components in the way described above and that provides the necessary forms and process facilities
- a Mover/Copier that takes care of URID database modifications (MPI already started building this component)

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

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

9.2 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 partner 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

- access handling is done by the originating[11] institution
- the access rights information is part of the URID database of the originating institute
- Shibboleth is used to exchange user information
- every partner will set up Shibboleth services
- Apache is used as entry point to handle HTTP requests, redirection tables are set up by the partners such that metadata is open, but that all resource requests are handled by an appropriately designed resource manager
- a prototype RM will be developed, partners are free to develop their own component, but have to adhere to reliability requirements and carry out careful tests
- the MPI will adapt ANNEX and LEXUS to have test beds for the web-application scenario
- the partners will discuss the requirements for access management (processes, rights, ...)
- a prototypical Authorization Database will be designed, that will be based on the requirements
- for the management of access issues a prototypical Authorization Handler will be developed, which will integrate those requirements that can be implemented given the constraints of the

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

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

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

Appendix C : Brainstorming Note about Federations

On types of collaboration, leading to federalism

Draft 0.9 David Nathan ELAR 19 Nov 05

Level A DECISIONS

1. Decision making. Decisions of varying types.
e.g. "Broadly speaking, a federation provides a common framework that links separate organisations (or, perhaps, parts of the same organisation) who want to collaborate or share information in a trusted manner. For example, if organisations A, B and C want to share information between them, they must at least decide, and agree on, the following: The information they want to share …" etc etc.
(from http://www.matu.ac.uk/federations_intro.html)
2. Identity and representation. Who are the organisations? Who speaks for them (to each other and beyond?)
3. Aims
4. Relationships. These catalyse and underlie development of collaboration. Varying importance at each level
5. Structures for co-ordinating 1-4

Level B OPERATIONS

6. Operational interdependence
7. A simple typology
peripheral, e.g. develop potential standards etc, e.g. OLAC, DELAMAN
utility, e.g. add capabilities, e.g. DAM-LR, DELAMAN
process, e.g. core functions or organisation, e.g. ELAR-OTA (implies level C)

Level C RESOURCES

8. Resource interchange
9. Distinguish whether resources or operations have overt financial cost/value
10. Need clear and binding agreements between parties
11. Legislative requirements and liabilities
12. Federation services.
This term is used by Dspace. It might include software and communications systems such as distributed access and storage. Distinguish two aspects of software services:
(a) support interoperability (→ level B)
(b) create a topology (i.e. "non-interoperability" beyond federal boundaries)
e.g. not only distributed ("federated") access management, but also ingestion, training, publishing, fundraising etc.

Level D RIGHTS AND OBLIGATIONS

13. "Common destiny". A body which self-declares both its constituency and the willingness of members to create, commit to, and be bound by an ongoing collective will, operations and destiny. Because collaborations involve more central functions, this level adds *rights* and *obligations* to the previous levels
Here, *institutionalised* relationships assume more importance.
I do not discuss here the (more obvious) issues of roles, committees, membership fees etc that might be involved if formal federation was sought.