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/