2. 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 01/01/2006 to 31/12/2006

Content

1 Activity Report

1.1 Progress Report

1.1.1 Summary of activities and major achievements

At the various project and technical meetings (London, Nijmegen) all items that are relevant for the success of DAM-LR were discussed in great detail. In particular at the technical meeting in November important steps could be made towards an integrated Grid domain. In general we can state with respect to technical aspects that the project is on a good time scale.

- All goals scheduled for 2006 could be achieved and all deliverables were created. However, some unexpected problems that are normal in technology projects had to be overcome so that in total more person capacity needed to be spent than expected.
- The prototypical solution at the MPI containing all relevant modules such as Metadata Harvesting, Handle System for resolving unique identifiers, LDAP for authenticating home users and guests and Shibboleth as middleware for distributed authentication and authorization was tested successfully. A Shibboleth link with Lund and INL could be established to show the distributed authentication in work. The link with SOAS needs to be accomplished in 2007. All work is based on certified servers although some tests were carried out with proxy certificates. The state of the work is described in D9.2 in more detail.
- With respect to metadata harvesting all partners integrated their resources into one domain. Additional metadata descriptions were integrated into the archives. Due to the problems with the local setup from SOAS only a first collection could be integrated.
- All partners continued in establishing the local archive infrastructures. The partners invested great efforts to carry out the archive formation (WP4) work and to consolidating their local archiving system to contain the essential pillars (WP5/6/7).
    - The MPI archive grew from about 15 TB to 25 TB which is an increase by about 66% in one year. It extended its tape library to store this amount of data which is partly due to the fact that the archive was opened to other projects. The archiving system was improved and was installed in a number of other places. Except versioning and an improved access request system which is to come in 2007 it has all essential features.
    - LUND has a full system which is available via the web and it is now based on a new powerful storage server framework. Additional corpora were added and much was invested in training local people - scientists as well as technologists. In addition the major steps were taken to setup a system for the distributed Grid scenario. This work will be finished in January to be ready to do a demonstration in February.
    - INL was also very active in completing their local archive setup and installed the basic Grid components successfully. Also at INL much work was investigated to prepare the inclusion of new resources for example from the STEVIN project. The local archive system was extended to deliver metadata, to use persistent IDs and to integrate with Shibboleth.
    - As indicated SOAS was faced with major problems when setting up their local system. The archive system that was very well designed, but the work suffered from the leave of the key software developer. He could only start the development work, but not finish it. Therefore, a new developer was hired, but yet the system was not stable enough to allow for mass production.
- The discussion about the nature of federations was continued. DAM-LR itself did not intend to establish a federation, however, it was necessary to study to study the administrational, organizational and legal aspects of a federation of archives. Recent discussions at the eScience conference in Amsterdam showed that one can make a distinction between a "social" and a "commercial" model both resulting in very different assumptions about the level of detail to be fixed in agreements. In general academic federations can be based on shallow sets of agreements. Initiatives such as DAM-LR need to link up with the emerging national identity federations, but need also to consider European law.
- Very important is to draw conclusions as early as possible from a project such as DAM-LR, since this will help to adjust the expectations and may help other initiatives.

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. To be mentioned are in particular

- much awareness creation at the LREC and eScience conferences and the associated workshops where the DAM-LR partners were the driving forces
- world-wide language archive formation discussions within the DELAMAN network which brings together the major archives about endangered languages and music of the whole world and where the DAM-LR partners are leading the discussions
- the regional archive installations (Iquitos, Kiel) and preparation discussions (Lima, Buenos Aires, Rio de Janeiro, Tbilisi, Kiel, Canberra, Windhoek)
- two flyers were created as a joint effort of the DAM-LR partners in early 2006 (see appendices E and F) and another one mentioning DAM-LR explicitly was just finished by MPI.

Summarizing, we can state that the consortium is on schedule, however, in 2007 more effort needs to be invested to integrate the SOAS archive.

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	6.1	1.00	0.10	0.10	7.3

Management and general meetings (where all participants are invited)

Date	**Title/Subject of meeting**	**Location**	**Number of attendees**	**Website address**
25-27.1	DAM-LR Training Course	Lund	40	
25.8.	DAM-LR Meeting	London	8	
29.8.	DAM-LR EC Meeting	Teleconference	4	
9-10.11	DAM-LR Technical Meeting	Nijmegen	9	
8.12.	DAM-LR EC Meeting	Teleconference	4	
10/13.10	Meeting with Schöntal	Nijmegen	3	
several	Meeting with Soddemann	Nijmegen	3	

Thomas Soddemann is the technical advisor within the DAM-LR project to assure that knowledge worked out for example in the DEISA project is reused and that DAM-LR is adhering to global standards and best practices.

Milestones and deliverable achievements

Deliverable / Milestone No	**Deliverable / Milestone Name**	**WP / Subtask No**	**Lead Contractor**	**Planned (in months)**	**Achieved (in months)**
1	Annual Report	1	MPI	T24	T25
1	Federation Note	1	MPI	-	T24
1	First Conclusions	1	MPI	-	T24
2	Running Prototype	2	MPI	T14	T14
3	Integrated Metadata Domain (SOAS to come)	3	Lund	T24	T24
4/5/6/7	Archive Systems	4/5/6/7	SOAS	T16/T18	T21
8	Definition Report	8	INL	T24	T24
9	Distributed Solution Prototype	9	MPI	T20	T22
12	Training Plan	9	Lund	T24	T24

1.1.3 Project activities

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)	6,1	1	0,1	0,1	7,3
Local Prototype (WP2)	24,5			2,5	27
Metadata Integration (WP3)	4	6		12	22
Archive Formation (WP4)	40,4	11	11	49,7	112,1
Local Lund (WP5)		5			5
Local SOAS (WP6)			3		3
Local INL (WP7)				8	8
Definitions (WP8)	1			3	4
Distributed Solution (WP9)	32,6	1	2	2,5	38,1
Dissemination (WP12)	4	6	1	1	12
Total (person months)	112,6	30	17,1	78,8	238,5

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.

1.1.4 Special Management Difficulties

In this chapter we want to give an overview about efforts that were taken by every partner to overcome the severe funding problems.

In the proposal to the EC it was stated that all partners invest large sums to setup and continue local archives and that this was done by funds from own funds, government, but also other partners such as in the case of MPI from the VolkswagenFoundation. All these funds were added to result in the total contribution to the DAM-LR project as described in the summary table that became also part of the Technical Annex. So, formally speaking we realized not earlier that in summer 06 that funds from third parties cannot be included. The reason was probably that no auditing was required to test the specifications in the 05 management report. Another error that was made for this SSA project is the expectation that no equipment should be bought, but the partners understood it so that no equipment should be bought from the requested contribution of the EC. This error based on misunderstandings could be solved by creating an amendment to the contract.

Since only 10% money of the EC was requested in this SSA it was obvious that this money was highly required for developing the Grid layer and could not be spent on management. This resulted in a management solution that was not appropriate. In summer 06 Peter Wittenburg took over the complete coordination which resulted in a visit at the EC headquarters in Brussels in September. Only this visit could clarify the misassumptions that were made and that were not noticed before.

The amendment now allows us to include equipment in the budget calculations, of course, it could not help in the point that funds from third agencies could not be included. This created severe problems for MPI and SOAS. At MPI we could not hire another person on the project yet, i.e., all work - the management, the setup of the local archive and the Grid integration - was made by own staff which was not expected. At SOAS it turned out that almost no funds could be requested from the EC anymore due to their funding scheme. Of course, this had severe consequences for their plans and also in this case people helping in setting up the local archive and getting the integration done could not be hired as expected. Much overhead was involved to discuss this at all layers, which obviously took more time than hoped. Yet the audit process will bring final clarifications.

Of course, the DAM-LR project decided that the DAM-LR plans should be realized as intended and should not be affected by these difficulties. Since some of the partners are experienced with standard

EC projects we can only conclude that the specifications of the specific rules of the SSA led to these misunderstandings. We need to thank the EC officers to help finding a proper solution.

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/T24	Annual Report	1	MPI	T24	T25
2	D2.2	Prototype Report	2	MPI	T14	T14
3	D3.1	Metadata Report	3	Lund	T24	T24
4/5/6/7	D4.1/5.1/6.1 /7.2	Archive Formation/Local Archives	4/5/6/7	SOAS	T16/T18	T21
8	D8.1	Definition Report	8	INL	T24	T24
9	D9.1	Distributed Solution Specification	9	MPI	T14	
9	D9.2	Distributed Solution Report	9	MPI	T20	T22
12	D12.1	Training Plan	12	Lund	T24	T24

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

1.3 Use and dissemination of knowledge

In its second year the DAM-LR partners increased their dissemination activities considerable. The following list gives an overview about the events where DAM-LR members either put the DAM-LR topics into the focus or at least mentioned the need of archive federations.

Date(s)	Who	Location	Number of attend.	Title, Website address
22-24.2.	DGFS Conference	Bielefeld	50	http://www.uni-bielefeld.de/dgfs2006/
13.2.	CLARIN Plenary Meeting	Paris	40	http://www.mpi.nl/clarin/
15.2.	IDS Conference	Mannheim	>100	http://www.ids-mannheim.de
17.5.	Grid Workshop	Nijmegen	9	
22.5	Research Infrastructures Workshop	Genoa	40	http://www.mpi.nl/lrec/2006/index.html
24-26.5.	LREC Conference	Genoa	>800	http://www.lrec-conf.org/lrec2006/
17.5.	Multimodality Workshop	Genoa	40	http://wwwhomes.uni-bielefeld.de/mmc06/
1.6.	DANS Opening	Den Haag	>200	http://www.dans.knaw.nl/nl/dans_symposia/1_juni/
15.-16.6	DOBES Workshop	Nijmegen	40	http://www.mpi.nl/DOBES/meetings/
19.6.	MPG GW Conference on Advanced Methods	Berlin	50	
22-23.6.	EMELD Conference	Michigan	40	http://linguistlist.org/emeld/workshop/2006/
8.8.	Seminar about Future Perspectives at BMBF Ministry	Bonn	30	
18-22.9.	Archiving Workshop	Iquitos/Peru	16	
25-26.9.	Archiving Workshop	Lima/Peru	40	
29.9.	Archiving Seminar	Buenos Aires	8	
6.10.	Ministerial Archiving Workshop	Rio de Janeiro	20	
30-31.10.	CLARIN Plenary Meeting	Budapest	50	http://www.mpi.nl/clarin/
2-3.11.	DELAMAN Conference	London	35	http://www.delaman.org/meeting2006.html

27-28.11	GWDG Long Term Archiving Conference	Göttingen	>50	http://www.gwdg.de/forschung/index.html
4-6.12	eScience Conference	Amsterdam	>400	http://www.escience-meeting.org/eScience2006/
4.12	eHumanities Workshop	Amsterdam	30	http://www.mpi.nl/clarin/eHumanitiesWorkshop2006.html
6.2.2006	”Apprentissage des langues premières et secondes”	Paris, Ministère de la Recherche	250	
Aug 3, 2007	"The New Order in Collection Development - Revisited" (org. by The Fiesole Collection Development Retreat)	Lund university		
Around the year	presenting DAM-LR as part of presentations of the Lund Humanities lab to research groups and delegations from all over the world;	Lund University, Humanities Lab	Total Nr visitors 2006 =appr 400	
March 2006	GURT Conference	Georgetown University, Washington DC	50	http://www.georgetown.edu/events/gurt/2006/program.htm
March 2006	SOAS Seminar	SOAS, London	25	http://www.hrelp.org/events/seminars/
May 2006	LREC conference	Genoa, Italy	60	http://www.lrec-conf.org/lrec2006/
June 2006	ELDP Training	SOAS, London	20	http://www.hrelp.org/events/workshops/eldp2006_6/
Sept 2006	ELAR-ELAP training	SOAS, London	10	http://www.hrelp.org/events/workshops/elap_elar2006/index.html
Nov 2006	DELAMAN Conference	SOAS, London	35	http://www.delaman.org/meeting2006.html
23.3	Dutch HTL Agency’s Day of the Corpora	Rotterdam, the Netherlands	30	http://www.tst.inl.nl
30-31.10	CLARIN Plenary Meeting	Budapest, Hungary	50	http://www.clarin.eu

A complete list of publications and papers that included DAM-LR matters is given as appendix D. The DAM-LR partners can claim that they spread awareness about the necessity of integrating language resource archives at the world level and that they reached a large fraction of discipline scholars. Many linguists and in particular computational linguists understood the messages that language resource need to be more accessible and that language archive federations implementing Grid technology are basic building blocks to achieve this.

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

At the management level the MPI had to resolve in close collaboration with the PO the problematic funding aspects that posed difficult situations for some partners (see above). Further, the information channels were maintained and the DAM-LR meetings were organized. A few workshops were organized as spin-offs of the DAM-LR work (LREC, eHumanities). All deliverables were produced without great delays. The financial reporting for 2005 had to be done again and the annual reporting for 2006 was carried out.

WP2/WP4

With the exception of versioning the development of the local prototype was completed to generate a full-blown archiving system that is ready for being integrated into archive federations supporting the essential pillars. In particular, the switch to LDAP for user management and authentication was not as simple as expected.

Also an extreme increase of the archive size (from about 15 TB to now about 25 TB) required a high productivity and carefulness of everyone involved. Currently, the MPI archive covers more than 250.000 archival objects that all have to be managed and to be made available via the web-site. The LAMUS archiving system was designed such that all uploading, integration and management work could be achieved.

We should mention here that the MPI archiving system has now been installed at some remote centers and that in 2007 other centers will be established. All these centers even being located in South American countries, Africa and Georgia are natural candidates for being included in a world wide archive federation.

WP3/WP8

For all uploaded sessions that have been fully processed metadata descriptions were created and integrated as well, i.e., we can assume that the number of metadata descriptions now also increased considerably compared to 2005. Also in this respect an enormous effort had to be done.

Due to the work to setup the distributed solution (WP9) some additional definitions such as the Shared Attribute Set needed to be agreed upon and to be documented.

WP9

The focus of the DAM-LR work was on developing and testing the distributed solution. At the MPI a complete setup for the distributed solution was realized that includes all relevant components (PKI, Certificates, Handle System, LDAP including a schema and the Shibboleth middleware. A special JAAS had to be configured as CMA for TOMCAT to contact the LDAP and ADS systems for authentification purposes. The detailed description is included in D9.2.

To verify the functioning of the local setup at the MPI tests were carried out with Self-Loops, i.e., the MPI site was used as resource provider and identity provider. Finally first tests were carried out between the MPI and Lund. The remaining tests between MPI, Lund and INL were left to January 2007.

WP12

The MPI team was very active in its dissemination policy to make as many linguists aware of the need to do language archiving to not loose our cultural memory and to increase language resource visibility and accessibility. This can be seen in the list of activities and publications.

1.4.2 Lund

WP1

Besides managing its own local efforts to form the archive, Lund was active in organizing training courses and writing deliverables. Further, it contributed to the discussion about the difficult financial aspects and to the reports.

WP3
Two new metadata reports were produced as updates of a first version (see deliverables). The archive in Lund was extended by some new corpora - all described with proper metadata.

WP4/5
The archive formation in Lund was intensified, since a complete new server was installed that includes a storage system with large capacity (49 TB) and redundancy mechanisms. The goal is to use this new storage system for material for many research groups in the arts and humanities faculty. Due to the training courses that were given to many researchers of various disciplines, we assume that the new storage infrastructure will be widely used and that indeed a center for digital cultural heritage will emerge. To support all this a technical staff was formed and trained. In particular, Thomas Schöntal visited the MPI twice to pick up all essential knowledge about various components.

The local archiving system was integrated into the Lund data services and its LDAP framework. The local system now is a full-fledged archiving system and therefore ready to be integrated into the DAM-LR infrastructure. To simplify the task Lund has chosen to follow the prototypical system from MPI as close as possible.

WP9
All components necessary to establish the DAM-LR integration were installed (PKI, Certificates, Handle System, Shibboleth) and integrated. First self-loop tests were carried out successfully and a first interactive test with the MPI as well. In 2007 the setup will be stabilized and optimized.

WP12
Lund was very active with respect to training humanities scholars to motivate them to use the new facilities and to include their valuable data. It seems that the DAM-LR and Life Archives ideas are accepted by an increasing number of scholars, although the realization of plans still takes much time. In addition, Lund experts participated in the aggressive dissemination policy of DAM-LR in 2007.

1.4.3 SOAS

WP1
The ELAR team at SOAS is working on a relatively small personnel budget and its focus is on digitising and collecting endangered languages material. As already mentioned a severe fluctuation in the development staff created severe continuity problems and much management effort was required to find a new software developer. Also a complete change in the administrational staff which created an additional complication of dealing with the financial difficulties mentioned above. A major contribution was given to the Live Archives flyer and the DAM-LR meeting was organized.

WP4/6
SOAS improved their digitization lines by setting up a Samba service to improve the audio workflow for the Dobbin workstation. Further the storage system was extended. New resources were added to the archive. A metadata set was specified, however, due to the change in the development staff it was not yet possible to come to a stable archiving system . For the archive a web-site was developed and the staff got a training in the matters of certificates and PKI system.

WP9
The SOAS team setup a number of essential DAM-LR components such as PKI, Certificates, Handle System and also installed a Shibboleth instance. Yet these components could not be integrated which will come in 2007.

WP12
Also SOAS participated in the aggressive dissemination policy of the DAM-LR partners. In particular it organized the annual DELAMAN meeting where archive federations were one of the relevant topics.

1.4.4 INL

WP1
In 2006, the INL continued to host and maintain the DAM-LR wiki website. Remco van Veenendaal took over the daily management of DAM-LR at the INL from Peter van der Kamp in April 2006.

WP2
In 2006, the INL finished work on a prototypical DAM-LR system. A first version of this system only worked on Linux and UNIX platforms, but later versions also worked under Windows (XP). The results of this work were made available to the other partners via the DAM-LR wiki. These results also sped up the work on the local and distributed solutions.

WP3
At present there are 2 (large) language resources in the INL DAM-LR archive: the Spoken Dutch Corpus (CGN) and the IFA Spoken Language Corpus (IFA). The CGN corpus consists of 900 hours spoken Dutch (12,780 audio files) with several annotation layers (about 120 GB). The IFA corpus was created specifically for research into phonetics and consists of 50,000 (Dutch) words spoken by 8 speakers with several annotation layers (about 22 GB). Next to the linguistic information, both corpora provide lots of metadata (e.g. about the speakers and the file formats) in IMDI metadata files. Next to containing the CGN and IFA corpus, the INL DAM-LR portal provides access to the Lund and MPI archives.
After an upgrade of the CGN corpus' metadata from IMDI 1.8 to IMDI 3.0 in 2005 the INL has obtained IMDI 3.0 metadata for the IFA corpus. The INL IMDI portal software has been adapted to be able to handle this latest version of the metadata.
Due to the use of the Handle system, an update of the IMDI 3.0 metadata was required: the metadata records need to store the(ir) Handles and the portal(s) must (at least) be able to present the Handles for referencing purposes. The metadata of the INL's language resources and the portal software have been updated to reflect these requirements.
After the beta version of the INL DAM-LR portal was made available online in the summer of 2006, some (meta)data "bugs" (e.g. broken links) were reported. As a result of these reports, some metadata repairs were made.
Once the DAM-LR infrastructure and portal(s) are up and running for all partners in 2007, the INL will add more language resources (such as text corpora, lexica and translation dictionaries) to the INL DAM-LR portal.

WP4/WP7
Having prepared a DAM-LR compliant archive setup in 2005, the INL continued to work on setting up and professionalizing services of the Dutch HLT Agency in 2006. One of the main activities in 2006 was the work on a “production line” for language resources[1].
As the Agency’s main tasks are to acquire, maintain, distribute and support language resources, the development of a production line system for all the Agency’s tasks/services was started: storage, backup, version, change and release management of language resources were being integrated into one system – based on the archive setup created for DAM-LR. The DAM-LR beta portal, available online via http://imdi.inl.nl is one of various solutions the Agency can now use for the distribution of their language resources. Other examples are offline copies (on CD-ROM of DVD) or paper versions (printed dictionaries). Software packages like the LAMUS archive management software and the ANNEX annotation exploration software can be integrated into this production line system.
More details on archive formation and the local INL solution can be found in the combined D4.1/5.1/6.1/7.1 Archive Formation and Local Lund/SOAS/INL Report.

WP8
In 2006, two versions (T18 and T24) of the Definition Report (D1) were produced in collaboration with the partners. The definitions in the report are based on the earlier versions of the report and updated with the agreements achieved at the training session in Lund, a project meeting at SOAS, a technical meeting at MPI and e-mail and telephone discussions.

WP9
The INL and MPI worked closely together on the development of (the first version(s) of) a distributed solution after both partners had created stable local solutions. The INL beta portal now does not only host the CGN and IFA corpora, but also provides access to the Lund and MPI archives. The work on

[1] Boekestein, M., Depoorter, G. and Veenendaal, R. van (2006). Functioning of the Centre for Dutch Language and Speech Technology. *Proceedings of the 5th International Conference of Language Resources.* Genoa : pp. 2303-2306. (http://www.inl.nl/images/stories/tstc/LREC2006/478_pdf.pdf)

the distributed solution culminated in a successful test of an interaction between the INL's and MPI's Shibboleth systems at a technical meeting at MPI in November 2006.

WP12
The main dissemination event for DAM-LR in 2006 was the LREC workshop (and following conference). Remco van Veenendaal was co-organizer of this workshop "Towards a Research Infrastructure for Language Resources" and the Dutch HLT Agency was one of the bronze sponsors of the LREC conference.
Other dissemination events in 2006 were the Agency's "Day of the Corpora" in Rotterdam (23 March, with 30 attendees), a CLARIN plenary meeting in Budapest (30 and 31 October, with 50 attendees) and an e-Humanities Workshop in Amsterdam (4 December, with 30 attendees). On various other events, the INL handed out the DAM-LR and Live Archive flyers[2] to promote the DAM-LR project.

1.5 Dissemination Plan

For 2007 the DAM-LR group will continue its active style of disseminating knowledge and plans to turn the experiences into a persistent strategy.

DAM-LR partners will

- participate in a number of conferences and meetings that are relevant in our field
- present DAM-LR ideas at meetings of ESFRI, UNESCO and similar organizations
- setup regional archives at various locations world-wide with the idea to extend the archive grid
- link up with national identity federations and emerging Grids such as in Finland, Netherlands, Germany, UK etc
- organize training workshops as explained in the Training Plan (D12.1)

To take care of continuity of investments, the DAM-LR partners will work on a pan-European extension to create a European wide Grid of language resource archives and will participate in plans to come to a persistent research infrastructure such as it is planned by CLARIN. For both initiatives a consolidation of the legal/administrational/organizational framework towards a formal "Federation of Archives" is important.

[2] These flyers are available for downloading from http://www.dam-lr.eu.

2 Management Report

Attached to this report are

- C-Forms of all partners
- the summary C-Form

3 Report on the distribution of the community financial contribution

Due to the problems in financial reporting no additional payment from the EC occurred in 2006.

Appendix A : Detailed Project Activities

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

Topic	WP	PNr
DAM-LR wiki hosting and maintenance	1	INL
Project administration and coordination	1	Lund
Preparation of DAM-LR related meetings	1	Lund
Preparation of Lund workshop and training	1	Lund
Co-writing Deliverables and Reports	1	Lund
Finish D3 Reports	1	Lund
Project administration and coordination	1	MPI
Project website maintenance	1	MPI
Maintain email list	1	MPI
Preparation of DAM-LR related meetings	1	MPI
Organization of DAM-LR related meetings	1	MPI
Preparation of Lund workshop and training	1	MPI
Preparation for LREC workshop and conference	1	MPI
Writing of D1.2 Annual Report 2006	1	MPI
Creation of DAM-LR Flyer	1	MPI
Creation of Live Archives Flyer	1	MPI
Creation of Live Archives Web-Site	1	MPI
Sorting out Financial Reporting Problems	1	MPI
Organization of DAM-LR meeting	1	SOAS
Contribution to Live Archives Flyer	1	SOAS
Finish D4.1 (16) Report (including 5/6/7.1)	1,4,5,6,7	MPI
Finish D2.1 Prototype Specification Report	1,2	MPI
Finish Local Prototype Final Report D2.2	1,2	MPI
Finish D3.1 (12) Report	1,3	MPI
Finish D8.1 (12/18) Reports	1,8	MPI
Finish D9.1 (14) Report	1,9	MPI
Finish D9.2 (20) Report	1,9	MPI
Local Prototype implementation, documentation and testing	2	INL
Implementation of Local Prototype (continuation)	2	MPI
Testing of Local Prototype	2	MPI
Build and Extend “local” Archive	2	MPI
Metadata Integration (conversion of CGN and IFA corpora to IMDI 3.0, metadata repairs and Handle incorporation)	3	INL
Optimize “Local” Metadata Infrastructure (continuation)	3	Lund
Add new Metadata	3	Lund
Optimize “Local” Metadata Infrastructure (continuation)	3	MPI
Add new Metadata	3	MPI
Archive Formation (“production line”)	4	INL
Build and Extend “local” Archive	4	Lund
Add new Resources to Archive	4	Lund
Acquiring and setting up equipment include 49 terabyte data provision server	4	Lund
Move “Local” Access Management Infrastructure to LDAP	4	MPI
Add new Resources to Archive	4	MPI

Set up Samba on server for Dobbin audio workflow	4	SOAS
SAN expansion	4	SOAS
Ingest new resources to archive	4	SOAS
Contribution to and editing of deliverable 4.1/5.1/6.1/7.1 (combined Archive Formation and Local Lund/SOAS/INL Report)	4,7	INL
Implementation of Prototype Solutions in Lund (continuation)	5	Lund
Testing of Prototype Solutions in Lund	5	Lund
Move “Local” Access Management Infrastructure to LDAP	5	Lund
Request Prefix from CNRI	5	Lund
Define Handle Postfix Syntax and create Postfix generator	5	Lund
Setup Handle Server at Lund and carry out tests	5	Lund
Adapt existing software to operate with URIDs	5	Lund
Request RA Status and create Certificates	5	Lund
Setup a PKI system	5	Lund
Suggest, Discuss and Agree on Shared Attribute Set	5	Lund
Catalogue software development	6	SOAS
Metadata development - ELAR set	6	SOAS
Archive website development	6	SOAS
Undertake RA training and RA manager status	6	SOAS
Local INL implementation, documentation and testing	7	INL
DAM-LR portal development	7	
Definition Report (T18, T24)	8	INL
Define an LDAP Schema	8	Lund
Implementation, documentation and testing of Distributed Solution (culminating in successful test of infrastructure with Shibboleth in November 2006 during technical meeting at MPI)	9	INL
Setup Shibboleth Instance and test it	9	Lund
Tests between Shibboleth Instances at Lund	9	Lund
Tests between Shibboleth Instances at Lund and MPI	9	Lund
Principle Discussion about Federation Agreement	9	MPI
Explain Architecture, Components and Steps	9	MPI
Give Tutorials about Components	9	MPI
Format for Authorization Records	9	MPI
Request Prefix from CNRI	9	MPI
Define Handle Postfix Syntax and create Postfix generator	9	MPI
Setup Handle Server at MPI and carry out tests	9	MPI
Adapt existing software to operate with URIDs	9	MPI
Build a Mover with Handle System Upgrade	9	MPI
Request RA Status and create Certificates	9	MPI
Setup a PKI system	9	MPI
Suggest, Discuss and Agree on Shared Attribute Set	9	MPI
Agree on Account Duration details	9	MPI
Define an LDAP Schema	9	MPI
Setup a Archive User Management LDAP System and Interface with NTDS/ADS	9	MPI
Extract "Local" Users to Archive User Management LDAP	9	MPI
Design Resource Request System	9	MPI
Setup Shibboleth Instance and test it	9	MPI
Design and Develop Authorization Management Tool	9	MPI

Configure JAAS as CMA for TOMCAT to contact LDAP and ADS	9	MPI
Tests with Shibboleth Self-Loop	9	MPI
Tests between Shibboleth Instances at MPI and INL	9	MPI
Set up and test Shibboleth	9	SOAS
Request prefix from CNRI	9	SOAS
Define Handle postfix syntax	9	SOAS
Setup Handle Server at SOAS	9	SOAS
Adapt and Integrate (based on the results of the Shibboleth test)	10	INL
Test (some early preparations)	11	INL
Dissemination	12	INL
Training (IMDI, Ingestion, Archive formation)	12	Lund
Giving talks and lab presentations	12	Lund
Articles and book chapters	12	Lund
Preparation for LREC workshop and conference	12	SOAS
Preparation for LREC conference	12	SOAS
Preparation for ELDP training	12	SOAS
Prepare paper for Delaman IV	12	SOAS
DAM-LR FLyer creation	12	all
Live Archives FLyer Creation	12	all
Regional Archive Flyer creation	12	MPI

Appendix B: Language Archive Federations

Basis for Federation Agreements[3]

Peter Wittenburg, Daan Broeder, David Nathan, Sven Strömqvist, Remco van Veenendaal
DAM-LR Project
25.12.2006

1. Introduction

The DAM-LR project can be seen as a test-bed for future infrastructures in the humanities and beyond to enable the eResearch paradigm. On a small scale it is a time limited project to establish a federation of language resource archives that share a Grid integration layer. It does so by testing out proper technologies that will allow us to do the required virtual integration. However, as a first paper about federation aspects (FFN) [1] pointed out, a federation is more than sharing a few technologies. It is also one of the tasks of DAM-LR to investigate these non-technical and primary aspects of "federations" especially in our domain. To prevent ending up in endless abstract discussions, however, we will do this based on the concrete needs of infrastructure projects such as DAM-LR and in particular CLARIN.

2. Types of Federations

When consulting Wikipedia for the term "federation" [2] we find the basic principles of state organizations which are the most "deep" domains where the term "federation" is used. We can read the following:

> A federation (Latin: *foedus*, covenant) is a union comprising a number of partially self-governing states or regions united by a central ("federal") government. In a federation, the self-governing status of the component states is typically constitutionally entrenched and may not be altered by a unilateral decision of the central government. The form of government or constitutional structure found in a federation is known as federalism (see also federalism as a political philosophy). It can be considered the opposite of another system, the unitary state

Such "Deep Federations" include detailed constitutional regulations that are ultimately broken down into legislative requirements that define constraints for example on the citizens. A federation is seen here as an alternative to a centrally organized state, since the members of such a federation retain some self-organizing power. There are many different examples for such federations that differ in the balance of power. The European Union can be seen as an example of a loose federation where the individual states retain much power and where the central government is comparatively weak.

In computational areas where very sensitive material is used such as in the medical domain, the virtual integration of data resources is also very much subject to very detailed regulations. So, for example, federations of hospitals have to establish an extensive set of rules of how to exchange and use patient data. In this case we can also speak about "Deep Federations".

Compared to such "Deep Federations" we can refer to a large number of "Shallow Federations" with much less detailed regulations. FFN speaks about the "Google-Federation" where participants restrict themselves to certain formats and principles so that their web content can be harvested to become indexed. There are even no explicit signed agreements, just a common understanding is sufficient to achieve world wide integration of open web content. All participants share the same common belief in the usefulness of world wide data mining, i.e., they share the same mission.

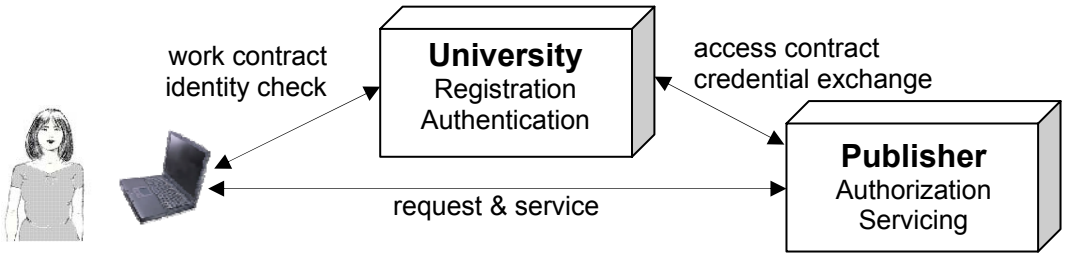
In the DSpace domain [3] users of the software discuss turning the user group into a federation that shares a number of interests that require a minimal governance rules such as enabling communication within DSpace community, ensuring the community is healthy and conflict resolution. Also in this case we can speak about a shared mission and a very loose definition of membership.

[3] This report is based on an earlier document created by David Nathan and Remco van Veenendaal which will be called First Federation Note (FFN).

N. Volanis and J. Dumortier [4] distinguish between two models of Grid computing[4] and describe their legal basis. The **social model** "views the benefits of grid computing as a resource to be harnessed for the good of the society". Meeting the social model's objective - the achievement of the scientific goal - relies heavily on the moral value of helping society by facilitating scientific research. The operational model depends on the voluntary submission of resources and in many cases the relationship between the partners is limited to the acceptance of terms of using given software. None of the actors engaged in the social model is willing to commit himself in a legally binding relationship that creates financial claims, obligations and responsibilities.

On the other hand, the **commercial model** sees in grid computing various business exploitation opportunities, i.e., companies need to control the resources to guarantee a Quality of Service. A number of enterprises can also form a Virtual Organization to share their data and resources based on a contractual relationship. These relationships will require severe financial constraints, controls and remedies, thus they require a "deep federation".

A kind of hybrid model is applied when for example large research institutions such as universities or groups of universities want to give their researchers access to a set of electronic versions of journals from publishers. The publishers will extend their normal set of regulations that define the usage of articles to the electronic domain and each user has to accept these rules. As can be seen in the following figure the university makes a contract with the publisher that gives persons with certain attributes such as staff member access to a number of eJournals. The researcher is contractually related to the university as staff member. When trying to access a paper the publisher will first ask the user to authenticate at the university so that some user attributes such as "is-staff-member" will be exchanged. Then the publisher will give access to the paper.



This dedicated federation is based on two contracts and trust that the university handles user attributes with care. The mission is well-defined for both sides: the university wants to give researchers access to all relevant publications and the publisher wants to ensure his income. The additional rules required by this Grid are comparatively shallow, since they only have to make specifications about the service to be delivered to certain members of the university, its technical implementation and the trust in the universities correct behavior. It may also make statements about the Quality of Service and specify penalties in case of misbehavior. This concrete model fits with the commercial model, however, in terms of our earlier discussion it is certainly a "shallow federation", since the number of additional rules will be small.

Summarizing, we can describe a number of characteristics that are typical for federations in the academic world:

- The partners share a **mission** that has to be made explicit and that every partner has to agree with.
- The partners have to describe the **trust relationship** which they all agree with, since in the strict sense their federations do not normally fall into the category "commercial model of grids".
- In general the partners in academic federations retain most of their **independence**, the federation just defines the regulations of the resource integration layer.
- The system of **regulations** is expected to be shallow, since topics such as quality of service are not an issue requiring severe penalties and since the ownership of resources will not be changed.

[4] The term "model of Grid Computing" is seen here as a synonym for a certain class of "federation models".

- **Penalty** regulations have to be defined in case of misuse, but since rights are not directly involved these can be kept simple. In general, exclusion from the federation will be sufficient which would require rules to decide this issue.
- Federations are not just made for a short period, but they add facilities at a structural level that have to be maintained with a **long-term perspective** to satisfy the needs of the researchers.
- According to Volanis and Dumortier this type of federation falls under the "**Information Society Services**" legal framework at least within Europe.
- A set of **technological agreements** have to be accepted by all partners to get the federation operational. Processes have to be defined how to maintain these agreements over the years and how to adapt them to new requirements.

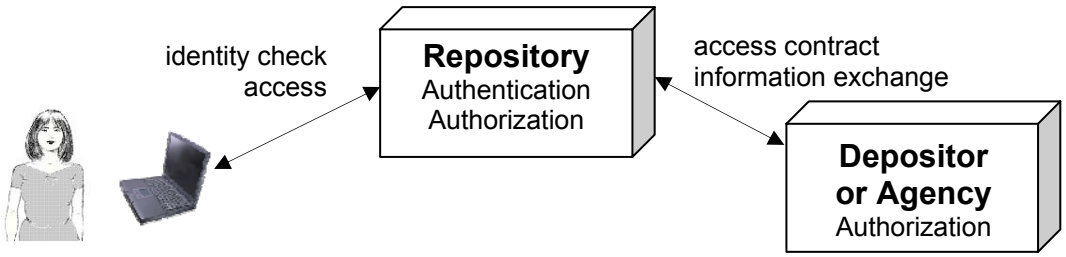
Federations in the academic domain turn out to be dynamic, i.e., new partners will join, others will stop their participation.

3. Rights Issue for the LR Domain

Essential to all regulational aspects in the language resource domain are issues that have to do with rights. To clarify the scope of the term "federation" we need some analyses of how a grid can influence the rights situation.

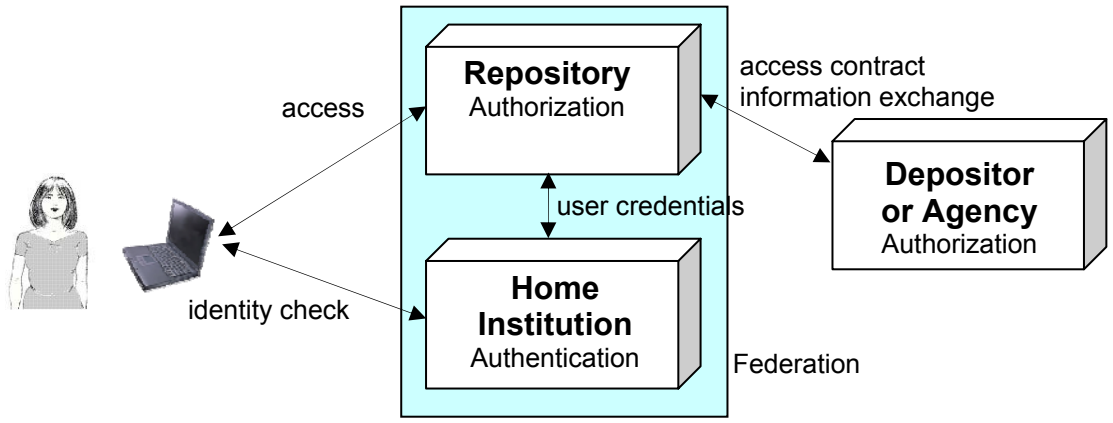
In general we have three important players when accessing language resources. We have the user who wants to access a certain resource that is stored in a repository. In general the resource is deposited by a researcher who has all rights in it[5] or it is provided by an agency that has these rights. In some cases the repository may have all rights in a given resource. Then the Repository also takes the role of the Depositor/Agency. In the following we will discuss a few scenarios where we will exclude the simple case that a resource is openly available via the web or where the resource is not accessible at all for anyone.

Scenario 1: This is the normal case where a user is dealing directly with the repository and where in some cases the repository will ask the rights holder whether access can be given. The repository takes full responsibility to handle access matters at a technical level as well.

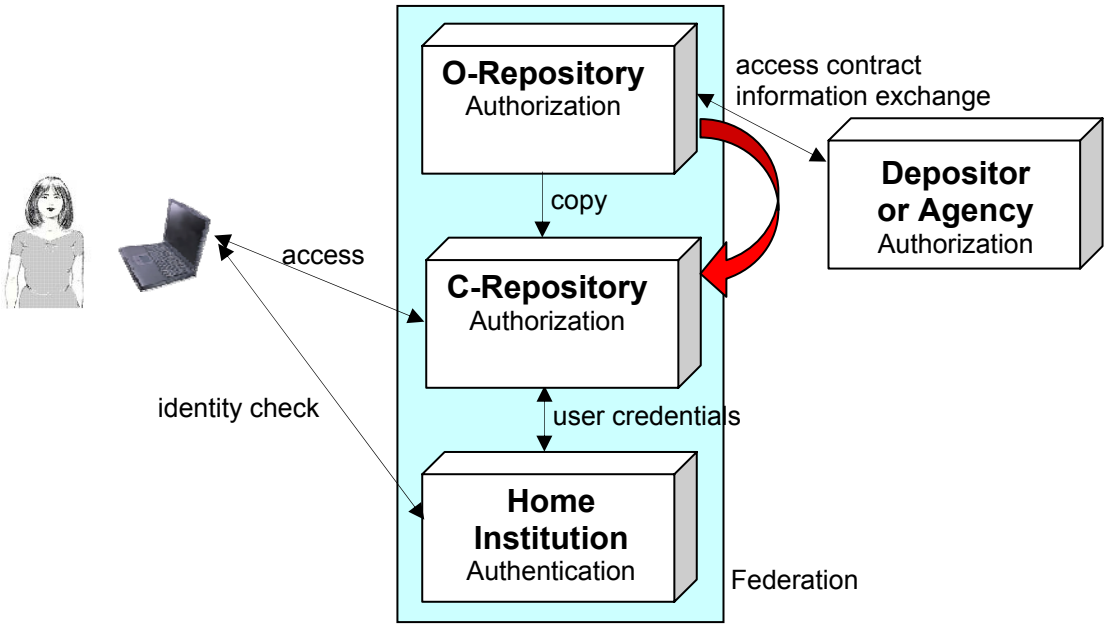


Scenario 2: In this scenario some additional components are introduced so that different instances form a "Shallow Federation". In the most simple case this just means that the functions "authentication" and "authorization" are split. A user that wants to access a resource has to first authenticate with his home institution which sends some agreed credentials to the repository, i.e., the repository relies on another instance to identify a user. The rights issues are not changed at all which makes federations of this sort very simple to establish. The trust relationship in the federation has to be specified, since we trust other archives to authenticate the right users, and give them access on the basis of this trust.

[5] We assume here that the repository has the right of archiving the data.

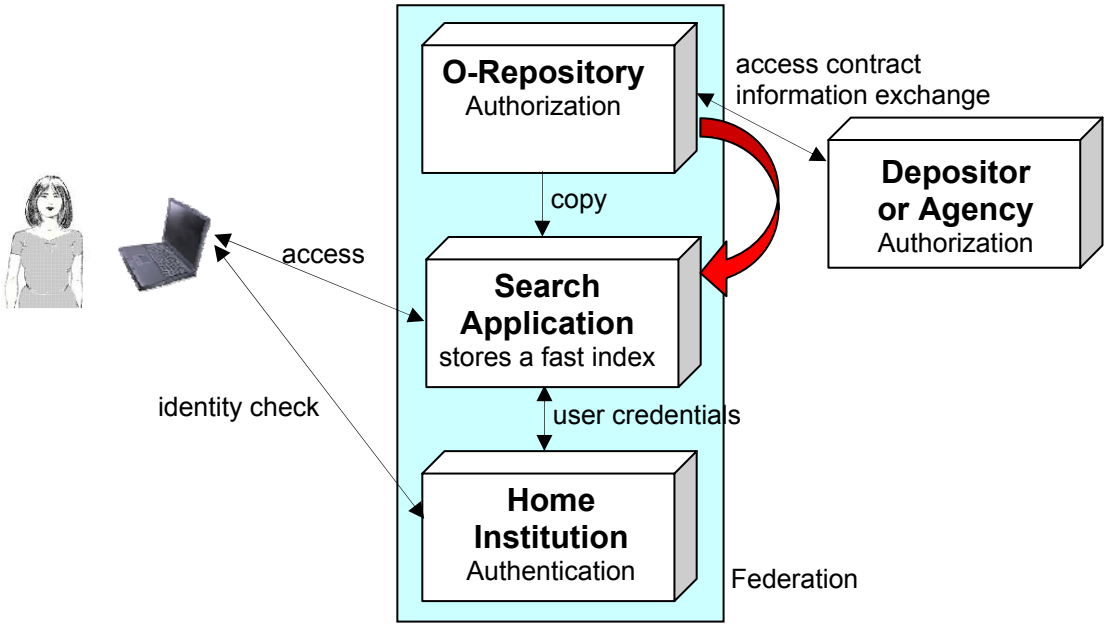


Scenario 3: In this scenario we assume that a resource is copied from the original repository to another instance which we call copy repository for several reasons such as long-term preservation and load distribution. This complicates the scenario slightly since the user does not interact anymore with the O-Repository that established the contracts with the depositor or agency, but with the C-Repository that does not have such a contract and probably even does not know any of the contract details.
The solution to solve this problem at a technical level is comparatively simple, since we only have to ensure that the rights on resources go with the copy and that the O-Repository (original copies) still is the only instance that may change them. Actually the technical solution would be different: the C-Repository would check at the O-Repository what the rights situation is and whether the requesting user is authorized to access a given resource. For this case the federation needs to be augmented with another trust relationship between the two repositories and probably some formal rules of behavior.



Scenario 4: We can think of a few mixed scenarios that can become very complicated to handle. These can emerge when for example applications are used that may be associated with graded access policies. As an example let's assume that a service provider wants to create an index about the contents of all resources in a number of repositories.
Of course, creating a fast index actually means copying the data and representing it in a different form that is optimal for searching processes for example. In the figure below one of the different possible architectures is given where the service provider running the search engine will receive a copy of all data to create the fast index, i.e., all data is copied to serve a new type of application. In other architectures the service provider would just receive the query and will send it in a formalized form to

the O-Repository that has its own fast search engine operating on the content. This does not imply a copy of the data, but nevertheless searching means to access the contents.



Probably, this type of access was not part of the contract between the O-Repository and the Depositor/Agency. This could be solved by amending the contract, but such operations are very costly and difficult, in particular, since there will be other type of applications as well. More simple is to stick with the former rule that any access to the content has to be granted according to the rights of the user launching the query. Technically this can be implemented by checking the access permissions for any resource that is accessed in the index or that results in a hit[6]. At the management level this can create a heavy load if there are no efficient management tools. Whatever the solution is the O-Repository has to rely on the proper operation of the application which requires a more careful consideration of the trust relationship and probably more complex regulations.
In the distributed case where the search engine is operating on the data at the O-Repository the responsible developers can implement all checks and algorithms that are required given the contracts with the depositors and they need not to rely on proper software from third parties. However, they need to invest in own software development that may be not feasible.

Summarizing we can say that a federation configuration does not per se make the rights situation more complicated, but that it introduces the need of new trust relationships. New types of services, however, can lead to rather complex situations.

4. Open Access

It is in the natural interest of researchers to have access to all digital resources that are available. In particular the web with its new possibilities allows to dream from a domain of digital resources free of barriers for the researchers. According to J. Taylor "e-Science [5] is about global collaboration in key areas of science and the next generation of infrastructure that will enable it". The his Cyber-Infrastructure NSF report the Atkins Committee [6] advocates for open platforms and referred to a Grid as an infrastructure for open scientific research. For specific domains (electronic publications) the e-IRG roadmap [7] even urges public funding for development of scientific software because current Intellectual Property Right solutions are not in the interest of science and the president of the MPG asks for new legal regulations that are not in complete opposition to current scientific usage scenarios enabled by modern communication methods and compliant to the framework of Open Access [8}.

[6] At the MPI one big index is generated covering all hosted resources. Including a resource in a query will only be given if the user has access rights for that resource. This seems to be a consequent and safe policy.

Data Grids are the kind of basic infrastructures currently being built up to create domains that integrate resources from different repositories, i.e., overcoming at least institutional boundaries to enable enhanced access and collaboration. However, there are still many obstacles to make resources openly available to researchers:

- There are and will be many resources that need to be protected due to privacy, religious and similar reasons, i.e., recorded persons don't want to be visible to the whole world.
- There are institutions that need to make some money to maintain their service, i.e., access needs to be controlled and some fee is required.
- The resources are partly donated from agencies that impose a restricted access policy and/or that want to get some money back.
- In "How open is e-Science" Paul David and colleagues [9] distinguish between e-Science and Open Science and discuss reasons for access restrictions that emerge from the research process itself.

Although many institutions fully support the Open Access initiative mainly as a counter movement to current trends of selling our cultural heritage to private institutions we need to realize that there are and will be many obstacles that will require access restrictions and sensitive access management policies. This is fundamental to our domain and any federation we create needs this both in technical and political sense.

Grid systems are being established to make access management feasible in the kind of distributed scenarios we are working on. When designed correctly they will not influence the legal situation between owners, resource providers and users, but simply require additional trust relationships.

5. Implications for DAM-LR

DAM-LR is a pilot project at small scale with a limited time span to try out suitable technologies to integrate language resource archives and to better understand the legal and ethical aspects involved. It is not an infrastructure project, although the participating institutions expect to maintain the integration layer beyond the project time. But there is no legal or financial binding of doing so. Therefore, DAM-LR follows the typical social model where the partners see the benefits of the integration for the researchers, indigenous people and other users.

The same is true for the world wide network of archives called DELAMAN [10]. With the help of the joined geographic representation as can be found at the Language-Sites web-site [11] one can easily find resources that are stored in the participating archives. Resources that were recorded from languages that are spoken in the geographic neighborhood can be found in close geographic locations. The map below gives an indication of this geographic integration despite all institutional boundaries.

This integration is done purely on the metadata level, i.e., this information is openly available anyhow. Accessing the resources themselves will always invoke the access managing components at the various repositories which is compliant to scenario 1 or in future also to scenario 2.

The DAM-LR partners agreed that at first instance scenario 2 will be implemented in all respects, however, its architectural design was made such that scenario 3 can be realized as well. It has to be seen whether a full blown copying scenario can be achieved or whether just a few tests will be carried out to show the proof of concept[7]. Also if there is time enough MPI will extend one of its web applications so that they can access resources from different repositories. However, the underlying access management model will still be as indicated in scenario 2.

Summarizing, we can say that DAM-LR just needs a shallow federation framework and that the partners need to agree on a number of technical aspects as being defined in the definitions documents.

[7] MPI is currently implementing software that can synchronize archives on the logical level of metadata organization and not just on physical level as it is for example done when only a backup for bit-stream preservation is done.

This Google Earth presentation of Language-Sites points to material from various languages stored at different locations such as DOBES archive (triangular icon), AILLA archive (salamander icon) and MPI archive (Greek symbol). This is just the beginning, since more archives have indicated their interest to join.

Technical Agreements

- every partner will offer his metadata descriptions as harvestable XML files to allow creating a unified domain of language resources based on the IMDI infrastructure[8]
- every partner will certify its servers and services according to the accepted TERENA TACAR list
- every partner will setup a PKI systems and sign its certificates with public keys
- the Handle System will be used to manage and resolve unique resource identifiers
- every partner is a Handle Authority and therefore has full authority to manage its own postfixes
- every partner will maintain its Handle System properly so that URIDs can be resolved
- MPI will setup mirror services for the different Handle Systems at the partner's sites, but will not modify any records
- access rights records will be associated with the URIDs in a unified format and managed only by the owning institution; these rights hold for all copies of resources
- every partner has to maintain and serve the agreed user attributes
- all authorization information for a certain resource is exclusively maintained by the originating institution - this right is not touched
- the access rights information is part of the URID database; it is up to every partner how this access rights information is maintained
- the choice for an authentication system is left to the partner institutions and a password identification mechanism is seen as sufficient
- Shibboleth is used to exchange user information, it is left to the partner institutions how they couple Shibboleth with their authentication solutions to extract the user information; however, the partners have to ensure that this interaction is operating correctly
- it is left to the partner institutes to decide about their resource manager, the component finally lending access to a resource[9]; it is the responsibility of the partners to solve the interaction between Shibboleth and the resource manager component

[8] The IMDI infrastructure is a result of EU funded projects such as ISLE and INTERA.

- DAM-LR (Shibboleth) differentiates between resource providers and identity providers, other institutes may be added as identity providers to allow their users to access resources from the partner archives, if they accept the rules and agreements
- every partner will provide mechanisms to request access permissions for a certain resource, it is left to the partners how they do access management

Federation Rules
With these technical agreements in mind and with the knowledge that only scenario2 will be implemented we can derive a few rules for the DAM-LR federation. Since in this scenario authorization is left with the institution where is was also in scenario1, only shallow rules have to be defined.

- the partners need to declare that they will handle all user information with care and that the user attributes are defined correctly
- the partners need to declare that access management definitions are left exclusively to the institution "owning" the corresponding resource
- the partners need to declare that they will maintain the required technical infrastructure as described by the technical agreements

For DAM-LR it is not necessary to have a certified portal that covers more than what is official for the DAM-LR project; in particular, there is no need for a unified set of rules describing correct behavior with respect to the usage of resources, since all access management aspects are handled by the "owning" institutions in the same way as until now. Such a portal could create, however, a kind of "federation identity". But this could also be achieved by using the DAM-LR logo on the own portal.

Since there are no structural funds reserved for the period after the end of the DAM-LR project, there is no legal binding for the partners, so we cannot expect a persistent infrastructure.

6. Grid Frameworks

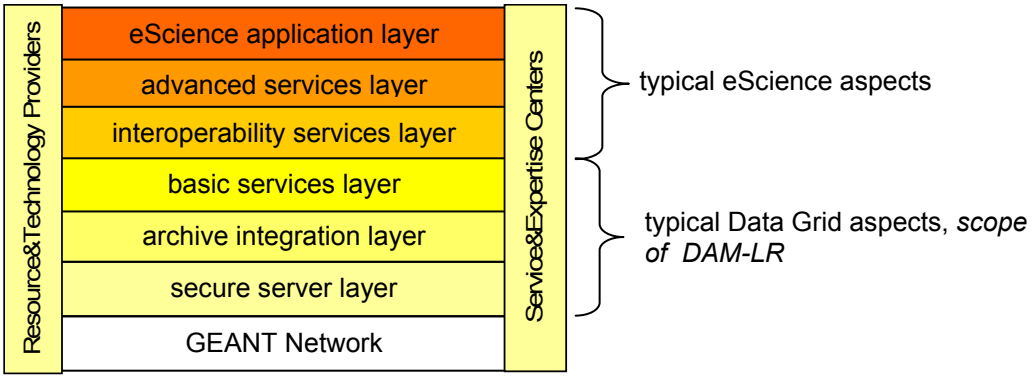
Grid systems are to a large extent discipline-independent, therefore we see different Grid infrastructures emerging at different levels: discipline crossing at European level (EGEE), discipline specific at European level (DEISA, DAM-LR), discipline crossing at national level (D-Grid), discipline specific at national level (Text Grid). All these initiatives are important to get a deep understanding about the service layers to be provided, the components to be used, the differences between the disciplines and the responsibilities at national and European level. One of the big challenges will be to bring all expertise and know-how together and to come to widely agreed standards to achieve a high degree of integration and interoperability. At this moment one can speak about highly coordinated activities at the national level in particular in the UK and in Germany as far as we know. This is due to the long-term intentions that are driven by the ministries and research agencies. At the European level the close interaction has still to happen and the need to harmonize things is enormous.

DAM-LR is a contribution from a specific data-oriented discipline in the humanities - the integration of language resource archives. Yet there are not so many experience gained in Grid projects in the humanities and social sciences (HSS) at a cross-national level. Even at national level there are not so many real Grid projects in the HSS. Therefore, the experiences gathered from DAM-LR are very important for further projects and research infrastructures at the European and at national level. Actually, the integration of repositories such as language resource archives only makes sense at a cross-national level, since only the virtual integration at cross-national level will create the new opportunities that researchers are looking for.

7. Implications for Research Infrastructures

Currently, the European Committee and the EU member states are discussing about building up persistent research infrastructures (RI) that will enable the e-Science scenario in various disciplines. Technically DAM-LR covers the basic integration layers of such a research infrastructure as indicated in the following figure.

[9] In most cases the Apache Web Server will play the role as resource managers.



It is obvious that RI have to go beyond what DAM-LR wants to achieve. Centers have to make formal commitments for a number of years with respect to the services they want to offer to the researcher community. They need to sign agreements where the even the Quality of Service (accessibility, availability) is specified, since otherwise the researchers can't rely on them. Here we can indicate a few such lower layer services that have to be guaranteed within a RI by a few centers for redundancy reasons:

- each federation needs to be augmented by official portals that contain certified and agreed documents
- several metadata portals need to be available to harvest metadata and exhibit a complete or split catalogue
- services to guarantee trusted servers and services are already maintained, i.e., all Grid projects can rely on them
- harvesting semantic mapping of metadata, metadata browsing and search services
- metadata schema registration service, metadata category registry maintenance
- Handel Services for institutes that can't run them themselves
- mirror services for URID resolving
- maintenance of formal lists (registries) of federation members and their characteristics
- it seems to make sense that a Europe-wide authority defines attributes for typical user types as they are accepted within the European research community, that all interested research institutions agree with the specifications and setup their user management so that these are maintained and can be extracted by authorized services; these specifications even should be synchronized at international level; they can widely be based on the specifications on the EduPerson agreements in the US
- certainly a European wide declaration will be needed that is signed by interested institutions that specifies that institutions will maintain user attributes correctly and that they will make them available to certified services in a federation; this general declaration will allow for example universities to join a merged resource domain without the need that for each Grid type of activity separate declarations have to be signed
- a list of accepted and certified components such as Shibboleth has to be maintained after extensive tests have been carried out; this guarantees that software components and applications that work with sensitive information such as user specifications are intensively tested before they can be applied in a federation
- of course many training services have to be offered
- ??

CLARIN Research Infrastructure

In the special case of the CLARIN infrastructure initiative [12] it will make sense to follow a two tier approach. On the one hand we should extend the DAM-LR Grid project to a pan-European dimension to build with some major goals in mind:

- extend all strategies to make them scalable,
- extend all mechanisms so that they can be turned over to a persistent research infrastructure,
- already establish the above mentioned services and give funds to the centers that will offer them,

- broaden the network of knowledge to the European wide community, carry out training courses and setup a help facility
- interact with other Grid activities at European and national levels to achieve unification
- prepare the Grid layer so that it can be integrated into a persistent research infrastructure for language resources and technology

On the other hand it is necessary to explore and accommodate the requirements for the much more complicated and to a large extent discipline specific higher layers to be tackled in a research infrastructure. Further, CLARIN can be the umbrella to work out all financial, organizational and legal aspects of a research infrastructure. A Grid project should not be loaded with these aspects.

References

[1] D. Nathan, R. van Veenendaal (2006). DAM-LR as a Language Archive Federation: strategies and prospects. LREC Workshop on Research Infrastructures, Genoa.
[2] http://en.wikipedia.org/wiki/Federation
[3] http://www.dspace.org/
[4] N. Volanis, J. Dumortier (2006). A European Legal Approach to Grid Computing. IEEE eScience Confererence, Amsterdam
[5] J. Taylor (2001). Presentation at e-Science Meeting by the Director of the Research Councils, Office of Science and Technology, UK, http://www.e-science.clrc.ac.uk
[6] D. Atkins, K. Droegmaier, S. Felman, et al (2003). Revolutionizing science and engineering through cyberinfrastructure. Technical Report, National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, D.C.: NSF
[7] M. Leenars (2005). e_infrastructures Roadmap: http://www.-e-irg.org/roadmap/eIRG-roadmap.pdf
[8] Open Access: http://www.soros.org/openaccess/
[9] P. David, M. den Besten, R. Schroeder (2006). How Open is e-Science? IEEE eScience Conference, Amsterdam
[10] http://www.delaman.org
[11] http://www.mpi.nl/services/mpi-archive/GE_language_sites
[12] http://www.clarin.eu

Appendix C: First Experiences from DAM-LR

- a technological note -
- first version -

Peter Wittenburg, Daan Broeder
DAM-LR Project
29.12.2006

1. Introduction

The DAM-LR project is one of the first Grid projects in the humanities - certainly at supra-national level[10]. Its task is to bring together archives with relevant language resources and allow users an integrated view on the resources they are interested in. On purpose, the four archives were chosen from different types of institutions and with different types of service focus. After having worked hard on the Grid implementation in 2006 we can draw first conclusions which will lead to a number of conclusions for the future.

2. Experiences

In this first version we present a number of points. It is not yet the moment to create a final coherent paper.

- The matter of integration is still a very complex task. Theoretically, it seems to be simple, but in praxis it offers many obstacles for the participating groups. Basically, this has to do with the usual design process, that needs to start with a restricted scope. Choices are made for concrete setups and solutions that turn out to be not optimal when it comes to an integration. Changes, however, are not always easy since they may affect workflow processes etc.
- We are lacking widely agreed standards relevant for our domain at many aspects such as for example metadata schema, user credentials, federation agreement types etc. Any expectation, however, that it just needs a group of persons who defines what the standard should be would fail, since we first need the experience from concrete projects.
- There is no off-the-shelf Grid technology, much relies on the availability of specialists who know about the details. Much adaptation and configuration work has to be done, which requires a deep understanding of the components. This is true even though the components used in DAM-LR (Apache, Tomcat, IMDI, LDAP, Handle System, Shibboleth) seem to be robust and reliable as expected. However, it is the interaction and integration that requires lots of efforts.
- On the one hand it seems that most the departments and institutions are not equipped with enough expertise to carry out the required work. On the other hand it seems that various computer centers have the required knowledge, but that the required experts are already heavily overloaded, so that they have to focus on certain projects, but cannot give services yet to all departments.
- Even in the case that an IT group is available, often the installation and integration cannot be carried out without expert help. This is due to the high work load of these groups, i.e., the potential experts that could be trained are sparse and overloaded with the normal tasks.
- We certainly lack a broad understanding at political level (university boards, institute directors, etc) about the general requirements put forward by these integration projects.
- The investments to establish such a Grid are considerable, yet there was no experience of how much efforts are needed to maintain a Grid. Through projects such as DAM-LR and others this becomes now more apparent. It is also obvious that the departments in general will not be able to maintain such a Grid with all its aspects over a longer period if there is no additional external expertise they can count on. The MPI is an exception in this respect.

[10] We will not discuss various national projects such as the D-Grid in Germany or the HAKA Infrastructure in Finland in detail. In particular the Finnish Grid is functioning system bringing together various universities and institutions including the humanities departments. In this system the Helsinki Computer Center takes the function of a national hub, but it only offers the Shibboleth distributed authentication.

- At the international level it became obvious that DAM-LR is ahead in the humanities compared to all other institutions. In the DELAMAN network for example archives from all over the world are collaborating (except Japan, China, Korea etc). From the DELAMAN meetings it is obvious that even our American colleagues cannot compete with us in this respect. Consequently, there is a great interest in joining our initiatives, i.e., the driving role of the Europeans is accepted. At our (MPI experts) South-America tour it became apparent that Brazil, Argentina and Peru want to establish central archive sites that participate in an international Grid of archives. They definitively want to join initiatives such as DAM-LR and get support even at ministerial level.

3. Conclusions

Although the DAM-LR project is not yet finished we can derive a number of conclusions that are important for future projects.

- Projects such as DAM-LR are absolutely necessary in particular in the humanities and social sciences to gather experience, but also to establish requirements, push standardization and broaden the awareness. Through our very active dissemination policy we have already created lots of awareness and interest in the emerging opportunities.
- The Grid requirements in the humanities and social sciences are different in many respects from those in the natural sciences. There is much more variation in the setup of the repositories, there is much more variation in the formats, the IPR issues are in general much more important, the chosen components are different etc.
- We certainly need more awareness and standards, but these have to be addressed at the European (if not international) level. Currently, too many Grid projects have a national scope. So at the EU level we need more interaction between the groups that are working and certainly need a body pushing forward standards. It would be extremely helpful if there would be official statements that certain protocols or agreements are compliant with the European rules. Yet the EGEE project does not have the outreach it should have. So it seems that a broader attempt is necessary.
- However, in many cases standards must not lead to severe restrictions. In the case of user credentials a EU wide unification would make sense for example, since this can easily be done. The same is true for unique resource identifiers and here we already have a kind of world-wide unification process going on. In the case of metadata, however, we are faced with a natural diversity that needs to be maintained. Therefore, a standardization can only be achieved through flexible component technologies and central concept registries as it is being established in the LIRICS project.
- Due to the fact that there is a lack of expertise it seems to be obvious that we will need a kind of specialized service group of people that have this expertise and the time to go to participating departments or institutions to setup the Grid system. Such a service group could also do remote monitoring of the status of services, something the MPI for example is currently already doing for remote archives. Such service groups should have a knowledge of the broader domain (such as humanities and social sciences) since they need specialist knowledge to make proper choices. Therefore, a European approach is the only one that makes sense instead of a national approach.
- Such an investment can be made within a project with limited time span, but we will also need a "research infrastructure" where structural money is available to give long-term support and to guarantee the long-term stability of the services. If there is no such guarantee researchers will not make use of it. In this respect the EC has taken the right strategic decision in time to spend funds on establishing such infrastructures.
- Time is ripe to broaden such initiatives and to built on the gathered experience. Scholars in the humanities are developing first research paradigms that need to built on Grid technology, it is time to create Grids of a critical mass in certain domains and to develop long-term strategies.

Appendix D: Publications & Presentations

Daan Broeder, Peter Wittenburg; Architecture and Components for an Archive Grid, DAM-LR Meeting, Lund, January

Daan Broeder, Peter WIttenburg. Language Archiving Technology. IDS Conference on Language Processing Technology. Mannheim, March

Peter Wittenburg. Towards a Research Infrastructure for Language Resources and Technology - an introduction. Research Infrastructure Workshop at the LREC Conference 2006, Genoa, May

Peter Wittenburg. Trends in Language Archiving Technology. COCOSDA Workshop at the LREC Conference 2006, Genoa, May

Peter Wittenburg, Daan Broeder, Technologies for an Archive Grid, Joint meeting MPI-DANS-BigGrid, Nijmegen, May

Peter Wittenburg, Language Archives – essential pillars for eHumanities, Official DANS Opening, Den Haag, June

Wolfgang Klein, Peter Wittenburg, Sprach Archivierungs Technologie, GW Conference about Computational Methods, Berlin, June

Daan Broeder, Peter Wittenburg, Methods for Distributed Authentication and Authorization, DAM-LR Meeting, London, September

Peter Wittenburg, Aspects of modern Language Archiving Technology, Linguistic Department, University of Lima, Peru, September

Peter Wittenburg, Aspects of modern Language Archiving Technology, CONICET, Buenos Aires, Argentina, October

Peter Wittenburg, Aspects of modern Language Archiving Technology, Ministerial Workshop on Language Preservation, Rio de Janeiro, Brazil, October

Daan Broeder, Andreas Claus, Freddy Offenga, Romuald Skiba, Paul Trilsbeek, Peter Wittenburg, Language Archiving at the MPI, 27th Annual DGFS Conference 2006, Bielefeld, February

Daan Broeder, Freddy Offenga, Peter Wittenburg, Peter van der Kamp, David Nathan, Sven Strömqvist. Integrated Services for the Language Resource Domain. Proceedings of the Research Infrastructures Workshop at LREC Conference 2006, Genoa, May

D. Broeder, P. Wittenburg. Technologies for a Federation of Language Resource Archives. Proceedings of the LREC Conference 2006, Genoa, May

Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson, Laurent Romary, Foundations of Modern Language Resource Archives. Proceedings of the LREC Conference 2006, Genoa, May

Peter Wittenburg, Daan Broeder, Peter Berck, Han Sloetjes, Alex Klassmann, Language Resource Archiving supporting Multimodality Research, Proceedings of the LREC Conference 2006, Genoa, May

Daan Broeder, Citing Archived Objects, The Fourth Annual Meeting of the Digital Endangered Languages and Musics Archive Network (November 2006).School of Oriental and African Studies. London, UK, November

Peter Wittenburg, Anforderungen zur Langzeit-Archivierung, GWDG Workshop zu Methoden der Langzeit-Archivierung, Göttingen, November

Peter Wittenburg, Distributed Access Management for Language Resources, GWDG Workshop zu Methoden der Langzeit-Archivierung, Göttingen, November

Daan Broeder, Remco van Veenendaal, David Nathan, Sven Stromqvist, A Grid of Language Resource Repositories, Second IEEE International Conference on e-Science and Grid Computing (e-Science'06) p. 132. AMsterdam, December

Daan Broeder, Peter Wittenburg, The IMDI metadata framework, its current application and future direction, International Journal of Metadata, Semantics and Ontologies 2006 - Vol. 1, No.2 pp. 119 – 132

Strömqvist, S., Breidegard, B. & Holmqvist, K. (in press). A new generation of infrastructure for research on basic language skills. Proceedings from the International conference *Apprentissage des langues premières et secondes* edited by M. Kail, M. Fayol and M. Hickmann. Paris, Ministère de la Recherche: Recherche et nouvelles technologies, 2006.

Strömqvist, S., Educating the Humanities for e-science - caring, sharing and creating added values. E-science conference, Amsterdam, Dec. 2006.

Nathan, David. 2006. Proficient, Permanent, or Pertinent: Aiming for Sustainability. In Sustainable data from Digital Sources: from creation to archive and back. Linda Barwick and Tom Honeyman (eds). Sydney, Sydney University Press. 57-68

Daan Broeder, Freddy Offenga, Peter Wittenburg, Peter van der Kamp, David Nathan, Sven Strömqvist. Integrated Services for the Language Resource Domain. Proceedings of the Research Infrastructures Workshop at LREC Conference 2006, Genoa, May

Nathan, David and Remco van Veenendaal. 2006. DAM-LR as a Language Archive Federation: strategies and prospects. . Paper presented at LREC Workshop: Towards a Research Infrastructure for Language Resources, 22. May 2006, Genoa, Italy
Munro, Robert. 2006 "Current design issues for digital archives: Architectures supporting value-adding access via a user's preferred language(s) and granularity of materials". Paper presented at the Georgetown University Roundtable on Linguistics, Georgetown University, March 2006
Nathan, David. 2006. Sound and Unsound Documentation: Questions about the roles of audio in language documentation. Paper presented at the Georgetown University Roundtable on Linguistics, Georgetown University, March 2006
David Nathan. 2006. Protocol and the language data lifecycle at ELAR. Linguistics Departmental seminar, SOAS March 2006
Broeder, D. , Veenendaal, R. van & Nathan, D. & Strömqvist, S. (2006). A Grid of Language Resource Repositories. 2nd IEEE International Conference on e-Science and Grid Computing.Amsterdam, the Netherlands
Veenendaal, R. van (2006). DAM-LR. Presentation at the Dutch HLT Agency's Day of the Corpora in Rotterdam, the Netherlands