Deliverable 4.1/5.1/6.1/7.1
Archive Formation Report
Local Lund/SOAS/INL Report

DAM-LR

011841

Distributed Access Management
for
Language Resources

implemented as
Specific Support Action

Contract Number: *011841*
Project Coordinator: Peter Wittenburg
Project Web-Site: www.mpi.nl/dam-lr/
Deliverable: D4.1/5.1/6.1/7.1
Authors: SOAS, INL, Lund, MPI
Responsible: INL
Date: 03.09.2006

Content

0. Preface - Merging of Deliverables D4/5/6/7

Due to the practical work it turned out that the activities focusing on archive formation (WP4) and local solutions (WP5/6/7) are so intertwined that it would lead to artificial reports if the partners would try to separate the different aspects. Therefore, the partners decided to merge these reports that give a detailed overview about the state of work at the different sites with the exception of the coordinating partner who described his related work in WP2.

This means that this deliverable contains statements about the local setup of the whole archive and how the various archives prepare their architectures to be ready for the DAM-LR integration. To document the latter part we include as appendix 3 the report from the recent London meeting of DAM-LR which was accepted in the meantime by the Executive Board.

Also at this meeting the final decision was taken to suggest a merge of the deliverables to the EU.

1. INL Archive Formation

1.1 Initiation/origins

The Institute for Dutch Lexicology (Instituut voor Nederlandse Lexicologie, INL) studies and collects Dutch words. These words are stored in a "database": the Language Database (Taalbank). The INL's TST-centrale (Dutch HLT Agency) maintains and distributes state-of-the-art digital (Dutch) Language Resources (LRs) such as the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) and the LRs stored in the INL's Taalbank.

The INL has been a member of many Language Resource projects. Several were international (European) projects, e.g. PAROLE, Telri, ELAN and Simple. Of these projects, the PAROLE and ELAN projects can be seen as important forerunners of the DAM-LR project: both worked towards uniformly accessible LRs. Also, while working with the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen on the creation of a management and upload system for language archives with changing contents (Language Archive Management and Upload System, LAMUS), the INL and MPI saw the potential of a unified web portal (with single sign-on) for access to and management of LRs: DAM-LR.

1.2 Aims/goals

The INL's DAM-LR goals are in similar lines to the general DAM-LR goals, as made publicly available via e.g. the DAM-LR Flyers (available for download from www.mpi.nl/dam-lr): to create an integrated and unified repository of distributed LRs, in order to enhance accessibility and usability for end-users (researchers, but also other interested parties). Sub-goals (or requirements) are the creation of trusted servers and services, deep metadata descriptions, stable and unique resource identifiers, user management and authentication, exchange of user credentials for access authorisation and the long-term potential for exchanging LRs to strengthen preservation purposes.

1.3 Funding

The INL is mainly funded by the Nederlandse Taalunie (NTU, Dutch Language Union) and to some extent by the University of Leiden. The NTU structurally finances the core activities of the INL; smaller and larger projects (e.g. the TST-centrale) are funded separately, either by the NTU or by third parties.

The following investments were made by INL in the first 18 months of the project:

- about 80,000 € was spent to setup the local computer, storage and network infrastructure
    - 48,000 € in 2005 (see DAM-LR annual report 2005)
    - about 30,000 € in the first six months of 2006
- (more than) 110 person months of its own staff were spent to setup and manage the new archive infrastructure, to co-develop the archiving software and to prepare its installation
    - 102 person months in 2005 (see DAM-LR annual report 2005)
    - the person months spent on DAM-LR related tasks in the first six months of 2006 are in similar lines to the 2005 investment

Of the (more than) 110 person months, 5.8 person months were funded by the EC in 2005. The DAM-LR annual report for 2006 will contain the INL's exact time

specification for 2006, but at the time of writing about 10 person months were invested in specific DAM-LR tasks (i.e. the time invested by the DAM-LR team).

1.4 Staffing

In 2006, a dedicated DAM-LR team of specialists was formed with complementary backgrounds: computational linguistics, project management, system administration and software development.

The following INL staff members form the core of the INL's DAM-LR team:

- project management: Jeannine Beeken, Peter van der Kamp/Remco van Veenendaal
- computational linguists: Michel Boekestein, Bob Boelhouwer, Remco van Veenendaal
- software developers: Vincent Wagelaar
- system administrators: Michel Longuich/Adel Sulaiman/Vincent Wagelaar

Remco van Veenendaal took over Peter van der Kamp's role of supervisor of the local technical work in May 2006. Adel Sulaiman temporarily replaced Michel Longuich as the INL's DAM-LR system administration in June 2006. Vincent Wagelaar became the DAM-LR system administrator and software developer/integrator in July 2006. From July 2006 onward the computational linguists have a smaller role in the DAM-LR project, because of the (more technical) work on Handle, certification, Shibboleth and – later – the test (scenario).

Next to the DAM-LR team, there are several other INL employees facilitating the DAM-LR work: the TST-centrale, EDP department, the financial department and the INL's management team.

A first version of a DAM-LR portal with the local INL archive was made available online in the summer of 2006. This portal has an integrated metadata domain with two large speech corpora (CGN and IFA), user authentication via (open)LDAP and runs on an Apache web server. The Handle system, certification and Shibboleth are being added to this setup at the time of writing.

1.5 Hardware architecture

The main part of the hardware architecture is a dedicated DAM-LR server with the following specifications: a 3.4 GHz single core Intel Pentium 4 with 1 Gb RAM and redundant 250 Gb disk storage. The operating system on this server is FreeBSD 5.4 (with Java 1.4.2).

This server will house the web server, (meta)data, certificates and portal software (IMDI web portal, Handle, LDAP, Shibboleth, etc.).

The INL has also recently acquired a backup server. The backup of the data on the DAM-LR server will be a main part of the INL's backup strategy.

The DAM-LR server will live behind the INL's firewall, but system administration will ensure accessibility from outside the INL.

1.6 Information architecture

The aim of the INL is to keep the information architecture simple and manageable. Therefore, an internal project has been started to create a production line for LRs. Input for this production line are the LRs maintained by and/or made available to the INL. The output can be any (converted) form of (any version of) a resource (or

the original LR itself), published on any of a range of media. This production line has been presented as a poster at LREC 2006: Boekestein, M. et al (2006), "Functioning of the Centre for Dutch Language and Speech Technology", in: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006 May 24-26, Genoa, Italy. Paris, European Language Resources Association.

Not all LRs within the INL are IMDI-based. Some of the materials, like written text corpora, use TEI or are stored in (relational) databases. TEI can be converted to IMDI[1] (e.g. via on the fly XSLT translation) and databases usually have export options (and the metadata of these materials is also available), so integration of these non-IMDI LRs in the DAM-LR portal is very well possible. INL still has to decide if these LRs will be converted on the fly or if they will be pre-converted into IMDI-based corpora.

Several LRs are already IMDI-based. These LRs, e.g. the Spoken Dutch Corpus and the IFA-corpus, will be used in the first versions of the DAM-LR portal at the INL. As soon as the mayor technical issues have been solved (Handle, Shibboleth), non-IMDI LRs will be (converted and) added to the portal.

The top of the IMDI hierarchy at the INL will only be a kind of folder with references to the individual LRs (and probably those of the DAM-LR partners). Because of the use of URIDs and IMDI metadata the virtual (online) browsing structure of the LRs does not have to reflect the exact physical (directory) structure. This enables the creation of a more user-friendly browsing structure than one that is directory-based. At the time of writing, the virtual structure present in the (IMDI metadata of the) CGN and IFA corpus are used in the INL portal.

1.7 Workflow systems

The LREC paper mentioned earlier presents a workflow or production line for LRs at the INL. In 2005, the INL and MPI have joined forces in the development of LAMUS (see annual report DAM-LR 2005), the Language Archive Management and Upload System. LAMUS enables the (access) management of LRs in a DAM-LR portal and also allows users to upload LRs themselves. As soon as the INL has a working DAM-LR portal with a significant number of LRs, the INL will start to integrate LAMUS into the portal.

Next to LAMUS, the INL and MPI have started talks about the use of ANNEX (Annotation Exploitation tool) by the INL. ANNEX allows users to browse and search online LRs using a user-friendly user interface.

1.8 Collection and accession policies

Most of the LRs maintained by the INL's Taalbank stem from government-funded projects. The Dutch Language Union (Nederlandse Taalunie) owns these LRs and also many of the LRs housed by the INL's TST-centrale. The TST-centrale project has been initiated to facilitate research in the area of language and speech technology: the TST-centrale maintains LRs, solves IPR issues and – in general – lowers the re-use threshold for LRs.

---

[1] DAM-LR partners are discussing the optimal (IMDI) metadata set for both written text corpora and corpora with multimedia components.

There is also a multi-million government-funded programme for the reinforcement of language and speech technology in the Netherlands and Flanders, STEVIN[2]. All results (LRs, software, etc.) of STEVIN will be made available via the TST-centrale. At the time of making these results available, there should be no remaining IPR issues that could interfere with their (re-)use.

As a result of the LRs having been created with public (tax) money, the LRs should be made available to users (mostly researchers) at low (or no) costs. Commercial organisations are expected to pay (more).

In general, there will be a two-layered access policy: the metadata is publicly available and the data itself is protected. Users who obtain access rights for an LR will be able to access all data, since there should be no IPR issues left that might e.g. require subset-based access. The DAM-LR team at the INL created a list of required access rights for the available LRs at the INL (in the near future): "Access rights at the INL". This document has been made available to the DAM-LR partners via the DAM-LR wiki website.

There may be one or two exceptions to the general rule (two-layered access): commercially available LRs and LRs under "open licences" (GPL, Creative Commons, etc.). The access policy for the first type of LR will need to be discussed with the commercial organisations – where the INL will have to be careful not to end up as an online shop for these organisations. Open licence LRs (e.g. IFA-corpus) pose no problem: they can be added to the DAM-LR portal with full (read) access for all users.

1.9 Data sources

The document "Access rights at the INL" describes the access rights for the available LRs at the INL (in the near future) and therefore provides a list of available LRs (in the near future). Below is an overview of the LRs that will probably be available at the INL (in the near future):

CGN: Data metadata	Speech corpus with annotations and IMDI
CGN: Documentation (evaluation) reports, etc.	Manuals, (technical) documentation,
CGN: COREX	Corpus exploration software
CGN: Tools	Tools used to create and validate the CGN
RBN	Reference list of Dutch (lexicon)
RBBN	Reference list of Flemish Dutch (lexicon)
ALVV: OMBI	Tool: Dictionary editor
ALVV: bilingual lexicons	Various bilingual lexicons
WNT: 1995	Official Dutch word list of 1995 (lexicon)
WNT: 2005	Official Dutch word list of 2005 (lexicon)
ONW	Old-Dutch dictionary
eLEX	Electronic lexicon (single word lexicon)
eLEX-m-lexicon	Electronic lexicon (multi word lexicon)

[2] For more information, please visit http://taalunieversum.org/taal/technologie/stevin/ (in Dutch)

INL-corpora: 5 million words: Data	Written text corpus
INL-corpora: 5 million words: Software	Corpus exploration software
INL-corpora: 27 million words: Data	Written text corpus
INL-corpora: 27 million words: Software	Corpus exploration software
INL-corpora: 38 million words: Data	Written text corpus
INL-corpora: 38 million words: Software	Corpus exploration software
Parole: 20 million words: Data	Written text corpus
Parole: Lexicon	Parole's lexicon
Parole: Software	Corpus exploration software
Neologisms section of ANW	Lexicon of Dutch neologisms
NL-Translex: Lexicons	Dutch – English, Dutch – French, French – Dutch, English - Dutch
NL-Translex: Software	Translation software for bilingual lexicons
NL-Translex: Corpora	Dutch, English and French text corpus
NL-Translex: Documentation	Manual, reports, etc.
Terminology lexicons	Lexicons
Term Extractor	Software tool for extraction of terminology from text
E-ANS	Electronic version of the Dutch grammar rules
Regional Dictionaries	Dictionaries
STEVIN project D-Coi	Written Dutch corpus
STEVIN project JASMIN-CGN	Spoken Dutch corpus
STEVIN project COREA	Co-reference tools and annotated text corpus
STEVIN project IRME	Multi-word database and software
STEVIN project AUTONOMATA	Spoken name database and software
STEVIN: LRs resulting from the 2nd and 3rd rounds	
STEVIN: Demonstrators	Various demonstrators (software)
LRs provided by / housed for third parties (not owned by NTU)	
Spell-check web service	Spell-check software
14th century text corpus	Text corpus
Eindhoven corpus	Text corpus (one of the very first)
IFA corpus	Spoken Dutch corpus

As can be seen from the list, the INL's LRs also include software and documentation. At the time of writing the idea is to make these LRs available via other means than DAM-LR (e.g. CVS for software source code[3]).

Many LRs have been created or will be created (STEVIN) using public (tax) money and are or will be owned by the Taalunie. The Taalunie is the main

[3] For more information, see e.g. http://en.wikipedia.org/wiki/Concurrent_Versions_System.

supplier of LRs, but there are some talks with other parties about distributing non-Taalunie LRs via the INL. All parties would benefit from being able to upload their LRs directly into the DAM-LR portal of the INL.

1.10 Ingest methods

At the time of writing, ingestion of data into the INL's DAM-LR portal is a manual process. The LR is made available to the INL, copied to one of the INL's servers or hard disks and then checked, converted to IMDI (if needed) and added to the portal.

As soon as the INL has a working DAM-LR portal with a significant number of LRs[4], the INL will start to integrate LAMUS into the portal. Using LAMUS, users will be able to ingest (upload) data directly into the portal, bypassing the manual process. The MPI has experience in installing the LAMUS software on non-MPI computers and the INL has worked on the development of the LAMUS system. We do not expect that integrating the LAMUS software in the INL's DAM-LR portal will be a problem.

1.11 Preservation policies and strategies

The INL has been creating and storing LRs for 30 years and has had to migrate data to newer formats several times. It is safe to say that there is a lot of experience on preservation of (electronic) data.

A very recent preservation activity was the purchase of a (new) backup server with a capacity of 2 TB. The LRs stored in the DAM-LR portal will be part of the INL's standard backup procedure.

Next to backing up LRs, the INL uses a CVS system to store all changes made to LRs, e.g. as a result of bug fixing activities. The changes must be stored and older versions must remain available. This may affect the DAM-LR URID (Handle) system, since all files have URIDs assigned. As soon as there are more than one version of a file, we need to ensure that the (new) URIDs point to the correct versions.

Another part of the normal preservation plan is the offline storage of LRs in a vault. If anything happens to a server all data must be restorable.

1.12 Format and naming standards

The metadata in the DAM-LR portal will be formatted as IMDI 3.0. The data "under" the metadata could in theory be anything; there is no limit to the formats accepted (yet). It may not be easy to start enforcing the use of certain standards, as the INL (not being a formal project partner) can only advise on the use of standards in e.g. STEVIN projects.

The good news is that many LRs made available via the INL use (open) standards: ASCII, XML (IMDI) and e.g. WAV. The LRs created by the INL usually have a TEI encoding, which is also a standard (and convertible to IMDI).

If conversion to a standard becomes necessary for proper ingestion in the DAM-LR portal, the INL can draw from experiences in (European) projects like PAROLE and ELAN or discuss these matters with the MPI, who have a strict standards-based ingestion policy.

---

[4] Ingestion of data is a secondary requirement. Having a portal with LRs is a primary objective.

1.13 User client groups

The primary users of the INL's LRs are researchers. In 2004, there were e.g. 569 registered users of the three large, linguistically enriched online text corpora. Probably the second category of users – in importance – is teachers. Teachers use the data mainly to find examples to support their lessons.

With the increase in internet use and the initiation of the TST-centrale, online use (browsing and searching) of the INL's LRs by "laymen" (people with no linguistic background) has increased. On the one hand there are more people surfing the internet, on the other hand the TST-centrale is trying to make as many of the INL's LRs as possible available for online browsing and searching (i.e. DAM-LR). This will have an impact on the interfaces the INL should make available to its users: we should not only take into account the researchers' wishes, but also those of the general public.

DAM-LR's main task is to see if it is possible to use state of the art technology to join LRs housed by the partners: DAM-LR is a technology and/or infrastructure provider. Creating a user-friendly interface for DAM-LR users will be an important development wish, but not or first and primary concern.

1.14 Dissemination policies and mechanisms

The INL has presented their DAM-LR membership and progress on various occasions and in various publications:

- LREC 2006: member of the organising committee on the pre-conference workshop on Research Infrastructures
- LREC 2006: co-writer of the paper "Technologies for a Federation of Language Resource Archives"
- LREC 2006: co-writer of the paper "DAM-LR as a Language Archive Federation: strategies and prospects"
- Presentation about DAM-LR on March 23 2006, during a one-day symposium on corpora, organised by the TST-centrale

Also, the INL frequently visits conferences like the CLIN[5] (Computational Linguistics in the Netherlands), uses third party newsletters for distributing news, contributes to the HLT magazine DIXIT of NOTaS[6] and hosts a website with e.g. an RSS news feed[7].

1.15 Legislative compliance

With the Taalunie as the main owner of LRs at the INL, Dutch law is the guideline for data protection, privacy and IPR issues. The INL works closely together with a Dutch law firm specialised in IPR issues[8].

1.16 Legal and contractual

In June and July of 2006, the TST-centrale, Taalunie and STEVIN have worked together to create licences for Taalunie-owned LRs. The driving force behind this work are the licences required for the use of the end-results of the STEVIN projects, but spin-off licences have been created for current LRs, like the CGN.

---

[5] For more information, please visit www.let.rug.nl/~vannoord/Clin/.
[6] For more information, please visit www.notas.nl.
[7] For more information, please visit www.tst.inl.nl.
[8] For more information, please visit www.klosmorelvosschaap.com.

These licences, together with the (similar) online licences the TST-centrale uses to grant users access to LRs, will become the basis for most contracts between the INL and LR users, including users of the DAM-LR portal.

At the time of writing, these licences are in Dutch only. As soon as English translations become available, they will be made available (as examples) to the DAM-LR partners.

1.17 Protocol policies and implementation

Since (at the time of writing) all the INL's LRs stem from or are owned by the NTU, the two-layered access policy mentioned at "collection and accession policies" will be implemented by INL in the first version of the portal. The infrastructure does however support a more fine-grained system of access; the first version of the system will simply not make full use of all possibilities. The document "Access rights at the INL" explains how the system works.

1.18 Financial

Using the metadata of the INL's DAM-LR portal will be free of charge. This information will be made publicly available. However, there is a fee for obtaining the data (as an offline copy). This fee is very low for non-commercial users, but businesses have to pay a considerable amount of money to get full access to LRs like the Spoken Dutch Corpus.

The INL has been investigating (and piloting) the use of (free) online licences, but this usually is for accessing (browsing and searching) subsets of LRs.

The INL will have to discuss with the DAM-LR partners how to map the current pricing scheme on the accessing of the DAM-LR portal.

1.19 Related services

Since the facilitating of re-use of LRs is one of the main tasks of the INL's TST-centrale, the TST-centrale's services include workshops and guest lectures on LR-related topics. E.g. there have been several CGN workshops. Some are tailor-made for a specific audience (researchers of teachers) and held at companies or universities (where the company or university provides the facilities) and some are organised at central locations in the Netherlands and Flanders and are focused at a more general audience.

At the time of writing, a new project is being set up at the INL to focus even more attention on the INL's LRs. It includes the introduction of (working with corpora like) the CGN in Dutch language classes in Dutch high schools and an online CGN training course.

2. ELAR Archive formation

2.1 Initiation/Origins

ELAR is one of three programmes comprising the Hans Rausing Endangered Languages Project (HRELP) at the School of Oriental and African Studies (SOAS), University of London. HRELP was an initiative of The Lisbet Rausing Charitable Fund (www.lisbetrausingcharitablefund.org/), first conceived in 2002. It consists of ELDP (funding programme), ELAP (academic programme), and ELAR (archive). For the archive, activities were begun when David Nathan was appointed in 2004 (as Director) and Robert Munro shortly thereafter as software developer. HRELP is currently the world's largest funder of endangered languages research.

2.2 Goals and policies

ELAR has two main goals:

- to provide preservation and access to digital endangered languages data
- to provide advice and services, including collaborative resource development, to those supporting endangered languages.

Policies include:

- ELAR holds only digital materials. In some cases we will digitise other materials in order to archive them
- ELAR's main client group consists of language documentation projects funded by ELDP; however, endangered languages materials may be deposited by any person
- ELAR does not insist on any particular content or formats. However, formats inimical to digital preservation are discouraged or refused
- ELAR follows the OAIS (Open Archives Information Systems http://nost.gsfc.nasa.gov/isoas/ref_model.html) terminology and architecture wherever possible

In recognition of the nature of endangered languages and their communities, ELAR pays particular attention to protocols - the collection, formulation, and implementation of restrictions and sensitivities associated with deposited materials

2.3 Funding

ELAR is primarily funded from a 28 million € grant from the Lisbet Rausing Charitable Fund for the operation of the Hans Rausing Endangered Languages Project over 10 years 2002-2012. The Archive share is approximately 1.4 million €. The archive also received and administered large funding from HEFCE (the Higher Education Fund for England) of 512,000 €, of which 476,000 €was spent on developing the archive premises, mass data storage, and computer and media equipment. In addition (and aside from DAM-LR), smaller amounts of funding have been raised for project work, totalling 20,600 €.

Funding summary:

The following investments were made by ELAR in the first 18 months of the project:

- about xx € was spent to set up the premises, mass data storage, workstations, specialist software,

    - 48,000 € in 2005 (see DAM-LR annual report 2005)
    - about 30,000 € in the first six months of 2006
- (more than) 110 person months of its own staff were spent to setup and manage the new archive infrastructure, to co-develop the archiving software and to prepare its installation
    - 102 person months in 2005 (see DAM-LR annual report 2005)
    - the person months spent on DAM-LR related tasks in the first six months of 2006 are in similar lines to the 2005 investment

Of the (more than) 110 person months, 5.8 person months were funded by the EC in 2005. The DAM-LR annual report for 2006 will contain the INL's exact time specification for 2006, but at the time of writing about 10 person months were invested in specific DAM-LR tasks (i.e. the time invested by the DAM-LR team).

2.4 Staffing

ELAR was originally planned to have one only staff member. However, in early 2004, a software developer post was created. In addition, a half-time digital technician post has been created, in recognition of the archive's role in providing advice and training about equipment etc.

Current staff:

- David Nathan, director
- David Evans, software developer
- Tom Castle, digital archive technician
- Chaithra Puttaswarmy, research assistant
- For systems programming, we engage a local specialist who is familiar with the SOAS network infrastructure on an ad hoc basis, for example to install MySQL and the Handle System on our LAH server (total cost approx 1,400 €).

2.5 Hardware Architecture

ELAR mass data storage is located in the main SOAS server room and consists of (a) a Dell PowerEdge 1850 server, named 'LAH', dual Xeon, 1 GB RAM with operating system Redhat Linux (b) Dell SAN - 2 trays of raid disks with online capacity of approx 8 TB, and (c) a Dell PowerVault 136T LTO2 tape library (nearline capacity 14 TB), all communicating by fibre-channel (optical fibre); total cost 81,000 €. Tapes are stored in ELAR's Kaso fireproof data safe, as well as sent to Oxford Text Archive where data is stored on the OUCS HFS system (we created a 5 year contract agreement with OUCS, at a cost of 28,000 €).

Locally we have 8 Dell and Macintosh workstations: two workstations (one Windows, one Macintosh G5) are dedicated to ingestion, conversions, and administration, and another Windows workstation devoted to Dobbin software (www.cube-tec.com/dobbin/), a specialist suite from cube-tec Germany that provides quality evaluation and workflow for audio materials. In addition there are another 6 workstations, some specialised, e.g. for video processing, available to clients to prepare their materials for archiving. Total cost (workstations, peripherals, Dobbin archiving system 71,400 €).

2.6 Information architecture and security

Data is stored according to OAI model (see above). Structure and filenames of deposits are recorded, then filenames are regularised and used as handles to

unique filenames stored in flat directories (one directory per deposit, named with mnemonic of depositor name). Metadata is stored in a custom MySQL database.

ELAR actually makes use of two servers. The main HRELP website (www.hrelp.org) server is externally provided through our contract with the commercial provider Blackfoot UK. Both the Blackfoot server and LAH (see Section 5) run password-protected MySQL databases. The Blackfoot-hosted database and website is concerned with administrative functions of the organisation, whereas LAH's database is dedicated to the storage and management of archival objects and metadata that will be accessed by DAM-LR portals.

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

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

For DAM-LR, our environment as two implications:

(i) we can cater for a very targeted user base through the existing MyHRELP system and its (current and future) registrants

(ii) MyHRELP will supply the user authentican, LDAP will not need to be implemented. However, we will need to investigate, together with DAM-LR partners, how MyHRELP can be made interoperable with Shibboleth. This is not expected to be too difficult, since MyHRELP uses standard and open-source components (see above).

2.7 Workflow systems

The default workflow is as follows: data, including media (audio, video) is received in digital form (digitisation is done on a case by case basis as an ELAR service and is not core archive workflow). The minimal deposit is a data file and a filled in deposit form (deposit form at www.hrelp.org/archive/depositors/depositform/index.html). Deposit of audio/video without transcription or annotation is discouraged. The deposit form provides deposit-global metadata, and the depositor can supply file-level or bundle-level (e.g. session) level metadata in standardised or structured formats. All data first goes to a "pre-queue" (if accession decision is pending) or queue (if accession is intended). Its components/properties are evaluated: metadata, text encoding and structuring, linking, audio quality, audio and video parameters, preservation properties. As a result, requests may be made to the depositor, or ELAR performs transformations/conversions/transcoding in consultation with the

depositor. Where binary proprietary file formats have been submitted (e.g. MS Excel), the data will be converted to a suitable preservation format; in some cases, the original binary file may be conditionally archived as well as a service to the depositor. At accession, the deposited data enters the archive collection, metadata is created, and data is made accessible where appropriate. Depositors can update metadata via the www throughout the lifetime of the deposit. These systems are currently being finalised and will be fully operational in October 2006.

2.8 Collection and accession policies

ELAR aims to archive quality digital resources for endangered languages. Accession to the collection is pending evaluation on the basis of language endangerment, content, and the data volumes, formats and structures. External parties may be consulted for advice about the status of a language or resource. Access restrictions formulated according to the deposit form will be implemented but lapse if not maintained. Media without symbolic annotation is discouraged. All video should be edited for relevance.

2.9 Data sources

Primary data source is output of projects funded by ELDP, the granting agency which is also part of the Hans Rausing Endangered Languages Project at SOAS. However, ELAR may accept any digital endangered languages materials, subject to available resources and evaluation of their quality and format.

2.10 Ingest methods

ELAR's data typically arrives as digital (and is increasingly "born digital"). Data transport is via optical disks or hard disks, often taking advantage of individual travel or contact. Currently, about 50% of data is arriving on CD/DVD, and 50% on hard disks.

2.11 Preservation policies and strategies

ELAR has two central goals: the preservation and the dissemination of endangered languages data. While digital convergence has made dissemination a function of several archives, ELAR's dissemination goals are primarily driven by the demands of the endangered languages area, where language data and expertise need to be delivered in a timely and appropriate way to support local language maintenance and revitalisation goals.

ELAR ensures data preservation though:

- a secure office environment
- enterprise class mass data storage equipment, comprising optically-linked RAID SAN arrays, robotic tape library, and server
- regular scheduled data backup
- backups stored at an alternative location in fireproof data safe
- disaster recovery copies to be stored at another institution
- co-archiving with Oxford University to ensure long-term data stability

In general our strategy will be based on using multiple repositories and data migration, not bytestream preservation. We believe that bytestream preservation is:

- a heavy drain on resources (to provide simulation software)
- counter to our aim of timely provision of usable resources for supporting language maintenance and revitalisation

2.12 Format and naming standards

ELAR recommends but is not prescriptive about accepted formats. ELAR recommends the following:

- text - plain text, with or without markup
- documents - plain text, PDF or postscript
- structured text - XML, other markup (with description of markup system)
- structured data in commonly available Office formats - ELAR will convert them to archive-suitable formats
- preferred character encoding is ASCII or Unicode; any other encodings to be clearly documented e.g. ISO 8859-5
- sound - WAV
- image - BMP, TIFF, JPEG
- video - MPEG2

Each deposit consists of files placed in a single directory (exceptions to be made when there are a large enough volume of files to affect retrieval performance)

Naming for deposits follows the conventions:

- 'FamilynameYearKeyword' where 'Familyname' is the (lower case) family name of the primary depositor, 'Year' is the 4 digit year of deposit, and 'Keyword' describes the deposit, typically the main subject language
- The directory name should not exceed 30 characters. Truncate the keyword if necessary.
- File names conform to the ISO 9660 standard (Level 1, but allowing lower case).

For persistent, unique, public identification of deposits, ELAR has installed and subscribed to the CNRI Handle System (http://www.handle.net/), which will allow for unique and persistent names across the DAM-LR federation of providers. ELAR has created its local metadata set, and will use MySQL and PHP via a mapping table to provide IMDI, OLAC and TEI metadata.

2.13 User client groups

ELAR recognises six client groups (in approximate order of resource priority):

- depositors
- language communities
- ELAP (Endangered Languages Academic Program) staff and postgraduate students
- academic linguistic community
- professionals associated with language activities, e.g. education authorities
- the interested public

2.14 Dissemination policies and mechanisms

ELAR has been active in disseminating and encouraging feedback on its archive architecture, such as in the following events:

- David Nathan. 2006. "Protocol and the language data life-cycle at ELAR". HRELP Seminar, SOAS
- Munro, Robert (with David Nathan). 2006. Current design issues for digital archives: architectures supporting value-adding access via a user's preferred language(s) and granularity of materials. *The Georgetown*

University Round Table on Languages and Linguistics (GURT 2006), Washington DC
- Munro, Robert. 2006. Multilingual metadata: between translation and coexistent annotations. *Proceedings of the 2006 ELAP Workshop on Meaning and Translation in Language Documentation documentation*. SOAS, London.
- Munro, Robert. 2005. The digital skills of language documentation. In Peter K. Austin (ed) *Language Documentation and Description Volume 3* . London: SOAS
- Munro, Robert and David Nathan. 2005. Introducing the ELAR information system architecture. The Third meeting of the Digital Endangered Languages and Music Archive Network (DELAMAN III), Austin
- Munro, Robert and David Nathan. 2005. Towards portability and interoperability for linguistic annotation and language-specific ontologies, *Proceedings of the E-MELD Workshop on Linguistic Ontologies and Data Categories for Language Resources (E-MELD 2005)* Boston.

In addition, ELAR has run international training events for grantees of the ELAP programme: see http://www.hrelp.org/events/workshops/eldp2006_6/, http://www.hrelp.org/events/workshops/eldp2005/. These workshops include as a major component training in preparing and archiving digital data.

For archive data, ELAR plans three main modes of dissemination:

- via a single point of access (SPA) portal created as a major goal of the DAM-LR project. ELAR is in the process of becoming a UK Registration Authority for Grid authentication certificates, which will allow those with physical access to SOAS to enable a means of authenticating themselves for access to DAM-LR resources
- via ELAR's LAH server (this server will also allow depositors to add and update metadata for their deposits)
- via the creation of value-added products from archive materials, in collaboration with depositors, to produce usable materials such as for pedagogy and language revitalisation.

ELAR does not acquire copyright in materials and access to materials will be on a strictly non-commercial-use basis.

2.15 Legislative compliance

ELAR operates under the UK Data Protection Act 1998, the Privacy and Electronic Communications (EC Directive) Regulations 2003, the UK Freedom of Information Act 2000, and SOAS ethics policy (see http://www.soas.ac.uk/research/index.cfm?navid=2414)

All accesses to data are conditional upon users (a) having relevant rights to the resource and (b) assenting to conditions of use (the assent is logged).

2.16 Legal and contractual

The ELAR deposit form creates the agreement between ELAR and depositors for the transfer and ongoing management of deposited materials. The deposit form is based on forms used by relevant international archives. It has been briefly

surveyed by lawyers, and has been workshopped by potential user groups as well as archivists.

2.17 Protocol policies and implementation

ELAR acknowledges that the endangered languages field inherently involves sensitivities and restrictions in regard to control of and access to data, and that those sensitivities and restrictions can change over time. Under the rubric 'protocol', ELAR has formulated a set of data management and access options that aim to satisfy depositors' and communities' needs as well as be feasible to implement and maintain. To summarise them: depositors (or their delegate), can authorise access to data by (a) anyone (b) nominated groups and individuals or (c) via specific request from a user to depositor; in the latter case depositors can opt to have requests sent directly from users or mediated via ELAR. For more details, see http://www.hrelp.org/archive/depositors/depositform/index.html. ELAR undertakes to implement protocol but also encourages depositors to make access as open as possible. For example, ELAR offers to anonymise texts, or where a portion of an audio or video is sensitive, we ask depositors to express restrictions on that portion only. Unless or until ELAR has methodology to suppress that portion, we can "fall back" to restricting access to the whole resource. In other words, observing protocol takes precedence over access.

2.18 Financial charges

ELAR does not charge for depositing or accessing data. Where delivering data requires producing and sending disks, charges may be made where volumes are significant, but not in the case of language communities who have no alternative means of accessing the data.

2.19 Related services

ELAR does not by default digitise materials but does so on a case by case basis where it is in the best interests of archive data quality. ELAR is keen to be involved in projects that develop resources from archive materials, especially those that can help strengthen endangered languages. ELAR has participated in several such projects, including exhibitions, CD-ROM production, and in-community language workshops.

ELAR also conducts training for ELDP grantees and in specific technical topics for ELAP postgraduates.

2.20 References

Munro, R. 2006. Systems Analysis and Design for Endangered Languages Archive (ELAR) http://www.hrelp.org/archive/design/archive_design.doc

Munro, R. and Nathan, D. 2006. Archive design for an OAIS compliant endangered languages archive.

3. Lund, Centre for Languages and literature

3.1 Initiation/origins

The Centre for Languages and literature at Lund University was inaugurated in 2004. It is a body of 10 language departments, a new research library and a new common resource for research and education, the Humanities laboratory. Research covers between 30 and 40 languages, from the major languages of Europe to small tribal languages in South East Asia. The Humanities laboratory is a cutting edge facility for the study of language behaviour online, such as speaking, reading, writing, gesturing. The laboratory has an eye-tracking unit, a phonetic-acoustic unit, a writing unit, a body-tracking unit, and an electrophysiological unit.

3.2 Aims/goals and policies

The aim of Humanities laboratory is to facilitate scientific and educational activities in the areas of language, culture, communication and cognition, by providing cutting edge equipment, competent staff to guide students and researchers in their usage of the lab, and by providing methods for the digitization, storage and metadata classification of research data (text, sound, picture).

3.3 Funding

The Centre for languages and literature, including the Humanities laboratory, was created by a combination of internal funding (Lund University) and external funding (The Crafoord Foundation, The Wallenberg Foundation). The total cost for the creation of the Humanities laboratory, the host environment for the Lund DAM-LR group, was around 5 million Euros. Lund University financed the physical building, the Wallenberg foundation financed the equipment, and the expert staff is financed jointly by Lund University and the Crafoord foundation.

In 2005 Lund university spent in total 138.000 € for personnel costs (23 pm) and about 75.000 for non-personal costs directly for the archive formation and local Lund setup. In 2006 we do not yet have a complete overview, but the new server has been ordered (about 400.000 €) and as can be seen from chapter 4 the work has been intensified.

3.4 Staffing

The staff of the Centre for languages and literature is around 300. Staff involved in the DAM-LR project include

- Kenneth Holmqvist, Associate Professor, co-director of the Humanities Lab
- Birgitta Lastow, head of technical group, Centre for languages and literature
- Johan Dahl, systems administrator
- Pierre Palm, systems administrator
- Thomas Schöntal, programmer
- Marcus Unesson, metadata manager
- Victoria Johansson, IT pedagogue
- Anne Börjesson, IT-pedagogue
- Catta Torhell, head of Library at Centre for Languages and literature
- Annakim Elmtén, IT-librarian
- Richard Andersson, PhD student, responsible for eyetracking metadata
- Per Henning Uppstad, PhD, responsible for keylogging metadata
- Damrong Tayanin, researcher, responsible for Kammu archive

- Mats Andrén, PhD student, responsible for the child language corpora
- Jordan Zlatev, Associate Professor, responsible for Thai corpus
- Gösta Bruce, Professor, co-responsible for SWEDIA corpus
- Susanne Schötz, PhD student, responsible for SWEDIA metadata
- Sven Strömqvist, professor, director of the Humanities Lab, coordinator of the Lund Dam-LR group

3.5 Hardware architecture

Between October 2005 and July 2006, the Lund Centre has made investments in major equipment totalling 5,925,000 SEK (646,380 Euro). The acquisition includes Eye-tracking equipment (733,500 SEK = 80,020 Euro), body-tracking equipment (823,100 SEK = 89,790 Euro), EEG/ERP equipment (736,270 SEK = Euro), and a 50TB server (3,632,000 SEK = 396,230 Euro) to house research data for DAM-LR distribution.

In January 2006, a language archive server was set up at the Lund Centre with the help of the MPI group. The new 50 TB server is currently being set up and will be operational late September or early October 2006. The architecture of the current metadata server will then be moved over to the new server and an authentication and authorisation system will be set up. This server will house the web server, (meta)data, certificates and portal software (IMDI web portal, Handle, LDAP, Shibboleth, etc.). Preparatory contacts have been taken with www.handle.net and hdladmin@cnri.reston.va.us.

3.6 Information architecture and security

The Lund language resources will all be IMDI-based. The security system to be implemented is Shibboleth.

3.7 Data sources

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

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

We aim at getting as much as possible of the above four content resources available on the server – with metadata, authorisation etc – during the autumn 2006.

3.8 Ingest methods

A lot of data are originally produced digital. However, large amounts of data relate to audio or video recordings (tape or cassette). The technical group in the Lund lab has built a hybrid machine for the digitisation of analogue signals from a variety of storage media (audio tape, audio cassette, video tape, DVD, floppy disk, laser disk, etc), to facilitate digitisation and content contribution to DAM-LR.

3.9 Dissemination and training

A first training event was organized in Lund January 2006 with some 30 participants from 6 countries. A second training event is planned to be held in the Lund lab in January 2007.

The Lund DAM-LR team has strengthened its cooperation with the Directorate of the Lund University Library, who is committed to enhancing the propagation of DAM-LR facilities to researchers using the library. The Directorate has hired a former IT-pedagogue of the Centre for Languages and literature, Anne Börjesson, to help realising this goal.

In addition, the DAM-LR concept has been explained to researchers and decision makers in various contexts, for example a network of French researchers on writing (Poitiers, July 2005), First international research workshop "Brain Mind Behaviour" (Lund, September 2005), International research workshop concerning reading and writing in real time (Lund, September 2005), Meeting with Wallenberg Global Learning Network incl. DAM-LR as a vehicle for research cooperation between Stanford and universities in Sweden (Stanford, March 2006 and Lund, May 2006).

The concept of distributed archive management as an extension of the modern research library and its potential for research and education was also explained in a book chapter published in late 2005 Strömqvist and Torhell, 2005).

3.10 Ethical issues

Ethical issues – in particular the integrity of subjects in scientific investigations - have recently been under much debate in Sweden. The policies and regulations are contradictory, as illustrated by a case in medicine at Göteborg University, where a researcher was urged to present his research data to the public. In short, it turns out that the rules and regulations of the national research council and associated ethical committees advocate very strict policies including, for example, the subject's right to ask the researcher to destroy the data at any point in time – a policy which would exclude distributing the data on the Internet. On the other hand, Swedish law and the so-called "Offentlighetsprincipen" (Principle of public access) dictates that any citizen has the right to demand to see research data at any point in time. In response to the situation, the Humanities Laboratory at Lund University has developed strict rules for the handling of research data produced in the lab. The data submitted to DAM-LR partly derive from the Lab and partly from other sources. We would welcome cooperation on how to formulate contracts concerning donors of data which originate from outside the lab.

3.11 Financial charges

The Centre for languages and literature is, as part of Lund University, a governmental organization, and the services associated with the Lund implementation of the DAM-LR framework are free of charges for the users.

3.12 References

Strömqvist, S. and Torhell, C. 2005. Språk- och litteraturcentrum – en plats för samverkan. In B. Nilsson, K. Ohrt and C. Torhell Från centralbibliotek till nätverk – Lunds Universitets bibliotek. Lund University: Library Directorate.

Appendix 1: Inventory of access rights for INL-LRs

INL DAM-LR workgroup

October 2005

Introduction

The Institute for Dutch Lexicology is one of the members of the DAM-LR[9] project. One of the tasks (work package 7) is to install a local solution for all four access management pillars[10] for the INL language resources (LRs). This document is the first step towards such a solution. It lists (the main categories of) INL LRs – our archive – and describes (our initial thoughts on) the access rights per resource (category).

The information in this document describes our specific situation, but some of it may be of relevance to other DAM-LR project members.

[9] Distributed Access Management for Language Resources. See http://www.mpi.nl/dam-lr
[10] Distributed Metadata, Unique Resource Identifiers, User and Group Management and Access Management

Access rights specification

A minimal set of access rights consists of read and write rights[11]. It is also useful to have entities to give these rights to: users and groups[12]. Abbreviations and a simple syntax are used to formalise permissions.

A summary of permission symbols used in this document:

U (User): individual users.

G (Group): groups of users

R (Read): read access

W (Write) write access

These symbols are used in the following syntax:

Data User|Group Right

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

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

All CGN data: U:RemcoVanVeenendaal R

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

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

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

Further thoughts on access rights

At some point in time (before accessing the language resources), users will have to sign a licence or otherwise request access. The user's name, password and access rights will then be added to some access management database, while the signed licence is stored somewhere safe. The procedure to obtain a licence and add users to the database is not part of this document[13]. Also, this document does not contain rules for deciding which user should be granted which level of access.

Users without a licence will have default access rights: they are considered members of the group Guests. The guest access rights are the lowest access level.

---

[11] At the time of writing, the INL does not yet support uploading of new data to our repository or changing of existing data. A system for managing and uploading of data to language archives (LAMUS) has been developed at (and in cooperation with) the Max Planck Institute, but not yet installed at the INL.

[12] This includes groups of groups.

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

All other access levels also have (inherit) guest access. At the lowest access level, users can access the metadata of all LRs[14]: all users have access to all metadata.

The access rights system described in this document does not distinguish between commercial, non-commercial users or any other type of user. Access has been granted (or not), that's all the system needs to know (see footnote 13). The main difference between commercial and non-commercial users usually lies in the requirements for obtaining a licence: commercial users usually pay (more).

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

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

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

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

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

All CGN data: U:132.229.188.0 – 132.229.188.100 R

or

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

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

INL LRs

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

- Corpora
- Lexica/dictionaries/word lists
- Tools

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

- Documentation

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

The Dutch Spoken Corpus

The Spoken Dutch Corpus project was aimed at the construction of a database of contemporary standard Dutch as spoken by adults in The Netherlands and Flanders. The intended size of the corpus was ten million words (about 1,000 hours of speech), two thirds of which would originate from the Netherlands and one third from Flanders. In version 1.0, the results are presented that have emerged from the project. The total number of words available here is nearly 9 million (800 hours of speech). Some 3.3 million words were collected in Flanders, well over 5.6 million in The Netherlands.

The corpus comprises a large number of samples of (recorded) spoken text. The entire corpus has been transcribed orthographically, while the transcripts have been linked to the speech files. The orthographic transcription was used as the starting-point for the lemmatization and part-of-speech tagging of the corpus. For a selection of one million words, a (verified) broad phonetic transcription has been produced, while for this part of the corpus also the alignment of the transcripts and the speech files has been verified at the word level. In addition, a selection of one million words has been annotated syntactically. Finally, for a more modest part of the corpus, approximately 250,000 words, a prosodic annotation is available.[15]

The CGN consists of data, documentation, COREX[16] and tools.

CGN: Data

In the table below, the group names and data categories are very detailed. Using (or slightly extending) this scheme, it would be possible to set access rights for the tiniest sets of data. At the time of writing, there are no licences for such specific sets of access rights. Users either have access or they have not[17]. But as we attempt to create a future proof access rights specification system, we thought to include an extreme example.

Group name	Data category	Access right
G:Guest	Metadata	R
G:Guest	Any data	-[18]
G:CgnFullLicence	All data	R
G:CgnAnnot	All annot data	R

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

G:CgnAnnotXml	All annot/xml data	R
G:CgnBptfon	All annot/xml/bpt-fon data	R
G:CgnBptfonCompa	All annot/xml/bpt-fon/comp-a data	R
G:CgnBptfonCompaNl	All annot/xml/bpt-fon/comp-a/nl data	R
G:CgnBptfonCompaVl	All annot/xml/bpt-fon/comp-a/vl data	R
G:CgnPri	All annot/xml/pri data	R
… (See CGN directories) …	… (See CGN directories) …	R
G:CgnTigCompoVl	All annot/xml/tig/comp-o/vl data	R
G:CgnAnnotText	All annot/text data	R
G:CgnAwd	All annot/text/awd data	R
G:CgnAwdCompa	All annot/text/awd/comp-a data	R
G:CgnAwdCompaNl	All annot/text/awd/comp-a/nl data	R
G:CgnAwdCompaVl	All annot/text/awd/comp-a/vl data	R
G:CgnFon	All annot/text/fon data	R
… (See CGN directories) …	… (See CGN directories) …	R
G:CgnWrdCompoVl	All annot/text/wrd/comp-o/vl data	R
G:CgnLexicon	All lexicon data	R
G:CgnFreqlists	All lexicon/freqlists data	R
G:CgnLexText	All lexicon/text data	R
G:CgnLexXml	All lexicon/xml data	R

Using the table above, user RemcoVanVeenendaal could be made a member of the CgnAnnotText and CgnLexicon groups to have access to the data in annot/text and to the lexicon files. All other data (except the metadata) would be inaccessible.

Granting people access to e.g. all Dutch .pri files is a direct extension of this scheme: users would be member of the groups CgnPriCompaNl through CgnPriCompoNl. Or the users could be made member of a new group AllCgnPri with groups CgnPriCompaNl through CgnPriCompoNl inside (which reflects normal system administration procedures and is more future proof). Predefining all possible groups is impossible; adding groups (and users) must be a feature of any software system implementing the system described in this document.

CGN: Documentation

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

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

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

CGN: COREX

COREX is the exploitation (browse and search) software for the CGN. The distribution and versioning of the source code is not an issue for DAM-LR: this will be dealt with outside this platform.

The binaries (executable version) of COREX can only be used with the CGN corpus (locally) available, so distributing COREX separately from the corpus makes no sense. If required, special versions of COREX, working on sub-sets of the CGN, can be developed and distributed, again outside the DAM-LR platform.

CGN: Tools

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

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

Corpora

The 5, 27 and 38 million words corpora

The INL manages three very similar corpora: the 5, 27 and 38 million words corpora

The 5, 27 and 38 million words corpora consist of one or more (ASCII) files with little text structure. Some metadata is available.

Another corpus is the Parole corpus, which is a collection of present-day Dutch texts, containing around 20 million words. The texts were obtained from various publishing houses and other third parties, which implied that their use was to be contractually defined (copyright). Use is permitted for non-commercial research purposes only, and access is restricted to rather small texts fragments, with proper reference of the source.) Some metadata is available.

The NL-Translex project resulted in three spin-off text corpora: Dutch, English and French. At the time of writing, the IPR issues have yet to be sorted out. We do include these corpora in our inventory, but it might be possible that we will not be allowed to grant anyone access.

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

Lexica/dictionaries/word lists

The lexical databases RBN (Dutch reference list), RBBN (Belgian-Dutch reference list), GB95 (word list of the Dutch language, 1995), GB05 (word list of the Dutch language, 2005), ONW (Old-Dutch dictionary), e-Lex (electronic lexicon) and the TST-m-lex (electronic multi-word lexicon) exist as single files (databases). The Parole project also had a (significant) lexicon as a deliverable. Although it might be possible to create finer-grained access rights sets, we only foresee “access or no access” for these files at the moment.

The neologisms list of the Common Dutch Dictionary (ANW) is implemented as a database and will be made available in the local solution as such: a single file with some metadata.

There are several bilingual lexicons resulting from the NL-Translex project: Dutch-English, Dutch-French, French-Dutch and English-Dutch. Each lexicon is an XML file. As with the NL-Translex corpus, IPR issues have yet to be sorted out.

Other files with bilingual data – LRs for translation purposes – are: Arabic–Dutch and vice versa, Danish–Dutch and Indonesian-Dutch. More bilingual files will be available in the future: Dutch-Estonian and vice versa, Dutch-Greek and vice versa, Dutch-Finnish and vice versa and Dutch-Turkish vice versa. These bilingual LRs stem from the ALVV (advisory committee for translation resources).

Group name (options)	Data category	Access right
G:Guest	Metadata	-
GGuest	Any lexicon/dictionary/word list	-
G:Rbn	RBN database	R
G:Rbbn	RBBN database	R
G:Gb95	GB95 database	R
G:Gb05	GB05 database	R
G:Onw	ONW database	R
G:Elex	E-LEX database	R
G:TstMLex	TST-M-LEX	R
G:ParoleLexicon	The Parole lexicon	R
G:NLTransLexDutchEnglish	The NL-Translex Dutch-English lexicon	R
G:NlTransLexDutchFrench	The NL-Translex Dutch-French lexicon	R
G:NlTransLexFrenchDucth	The NL-Translex French-Dutch lexicon	R
G:NlTransLexEnglishDutch	The NL-Translex English-Dutch lexicon	R
G:AlvvArabicDutch	Arabic–Dutch–Arabic data	R
G:AlvvDanishDutch	Danish–Dutch data	R
G:AlvvIndonesianDutch	Indonesian–Dutch data	R
G:AlvvDutchEstonianDutch	Dutch-Estonian-Dutch data	R
G:AlvvDutchGreekDutch	Dutch-Greek-Dutch data	R

G:AlvvDutchFinnishDutch	Dutch-Finnish-Dutch data	R
G:AlvvDutchTurkishDutch	Dutch-Turkish-Dutch data	R
G:AnwNeologisms	Neologisms data of ANW	R

The creation and distribution of "tailor-made" versions (subsets) of these files may become an issue for DAM-LR. See footnote 17 on page 26.

Tools

Many projects did not only create a LR, but also delivered (exploitation) tools. Some of these tools will/can be included in the local solution. Others are too system-dependant or outdated. Please note that at the time of writing, we only foresee the distribution of binaries via the local solution. The distribution and versioning of source code will be dealt with elsewhere[19].

OMBI (reversing tool for bilingual dictionaries) and documentation could be included as summarised in the table below.

Indexing and retrieval software was created for the 5, 27 and 38 million words corpora. This software was written in Vax Pascal. The retrieval software (with UI) was written using the SGM (screen management) library. It is highly unlikely that users would like to reuse these resources. Also, making the binaries available in the DAM-LR portal seems pointless as they are too system-specific.

The exploitation software for the Parole corpus mainly consists of Perl scripts (for security) on the server side and JavaScript on the client side. The (knowledge of building the) current Parole website could be (re-)used in the DAM-LR project (why invent the wheel again) with minimal effort (since the INL developed the software), but there are better ways of giving users access to the source code than via the local solution.

Finally, although the INL has a licence to use the NL-Translex machine translation system, it is not known when or if the binaries and/or the source code are going to be part of our LR archive.

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

Documentation

NL-Translex: Documentation

Documentation, like user manuals, of the NL-Translex project and other projects is available. It may be possible that all (user) documentation is made publicly available (R for Guest). Another possibility is that some documentation

[19] E.g. CVS (concurrent versions system) or SVN (subversion: http://subversion.tigris.org/).

(evaluation reports) is made available via a separate group (G:ProjectNameReports R).

An example[20]:

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

Future language resources

In addition to the LRs mentioned earlier, the INL will manage more LRs in the (near) future. Not only as a result of being a central repository for LRs resulting from government-funded projects, but also because other parties are starting to see the benefits of outsourcing maintenance, management and distribution work to the INL.

Some examples of possible future LRs are:

- Demo versions of LRs. These should be made publicly available (G:Guest R)
- Terminological lexicons of CoTerm (committee for terminology).
- Terminology extractor of CoTerm. This tool (currently under development) will fit nicely in the tools section.
- E-ANS (electronic version of the Dutch grammar rules)
- Regional dictionaries
- Results of STEVIN[21] projects
    - D-Coi: a pilot project for a 500 million words text corpus
    - JASMIN-CGN: extension of the CGN with speech of elderly, children and non-native speakers.
    - COREA: coreference resolution for extracting answers
    - IRME: identification and representation of multi-word expressions
    - AUTONOMATA: grapheme-to-phoneme LRs for proper nouns
- A corpus of 14th century Dutch text
- The Eindhoven corpus (or corpus Uit den Boogaart)
- Spelling web service. The spelling software will be a service available on the INL website. It is unclear is the source code (or data) will be made available via the local solution.

Supplementary thoughts

In the past year, the INL has developed a demo version of the CGN corpus (see Future language resources. This demo contains a subset of the data of the entire corpus. Since it is a demo version and publicly available, it might be possible to

[20] Most LRs have some kind of documentation, not just NL-Translex.

[21] Essential electronic language and speech resources for Dutch: http://taalunieversum.org/taal/technologie/stevin/. At the time of writing, only the first STEVIN call has been issued.

make this demo guest-readable. This would present visitors of the local solution with the opportunity to preview and/or evaluate the CGN corpus.

All materials without metadata should have metadata created[22] to make browsing and searching the online catalogue of LRs easier.

Conclusion

This first step towards a local solution for the INL LRs offers a simple, but practical description of a system of access rights that can be used to implement a tool for managing user access to the INL resources. We have presented examples of how the scheme can be applied to small subsets or entire collections of the LRs and we have given some insight into the many (types of) language resources at the INL.

Armed with the information in this document, we will investigate if we can use available tools to implement a local solution or if we have to create a solution from scratch.

[22] IMDI metadata will be created with the IMDI Editor: http://www.mpi.nl/IMDI/tools

Appendix 2: Archive Data Models at SOAS

Final / Physical Data Model

Final dataflow diagram

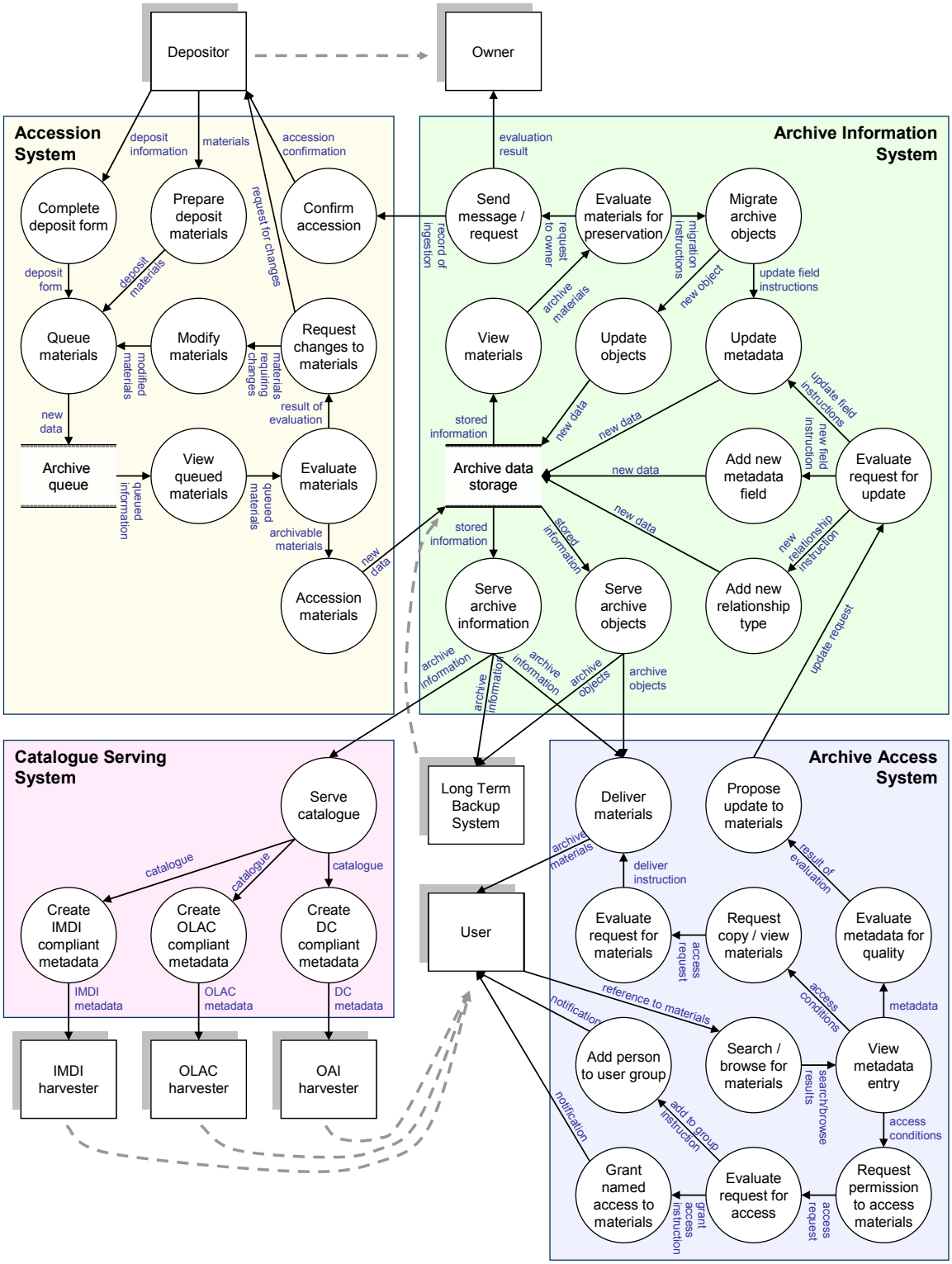


Figure 3.1: Final combined Data Flow Diagram

Combined Physical Model

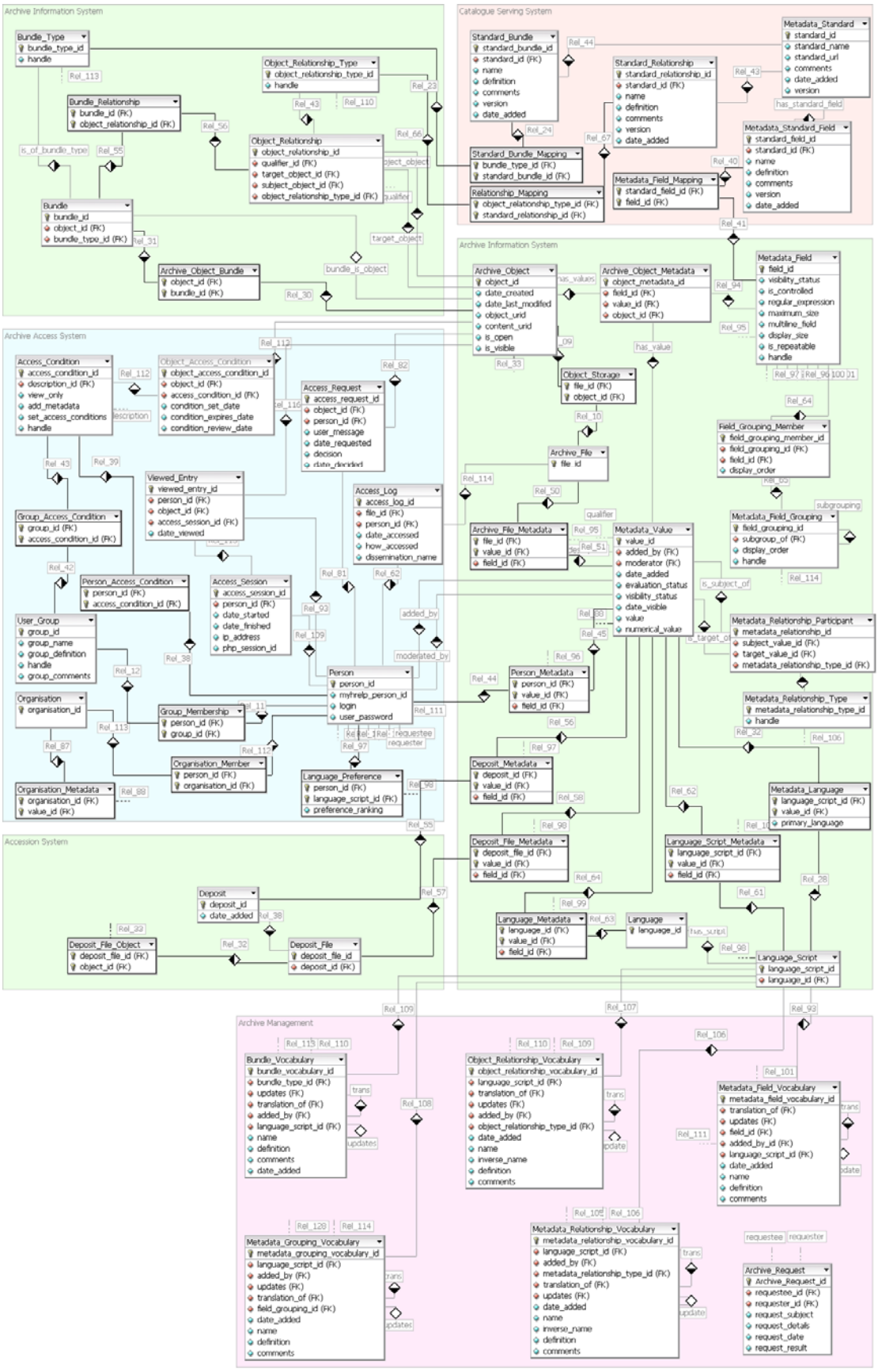


Figure 3.2: Combined physical model of the four systems.

Accession system

Data Flow Diagram

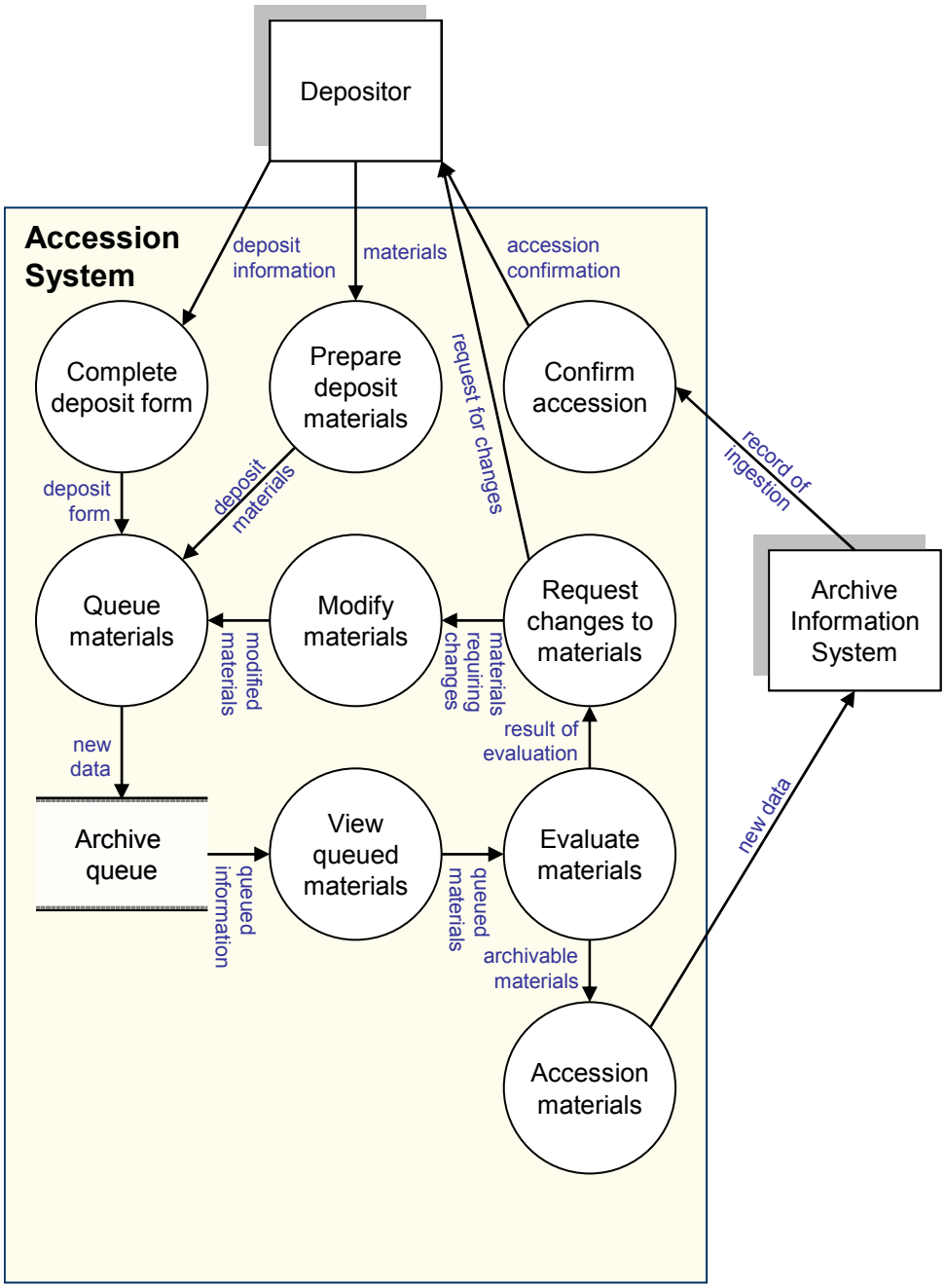


Figure 3.3: Final Data Flow Diagram for the accession system

Physical Model

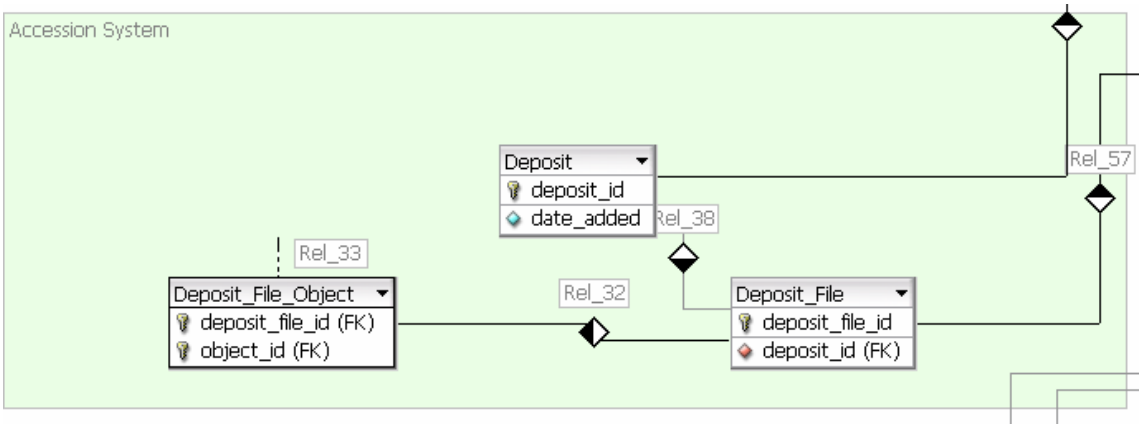


Figure 3.4 Physical model for the accession system

Archive information system

Data Flow Diagram

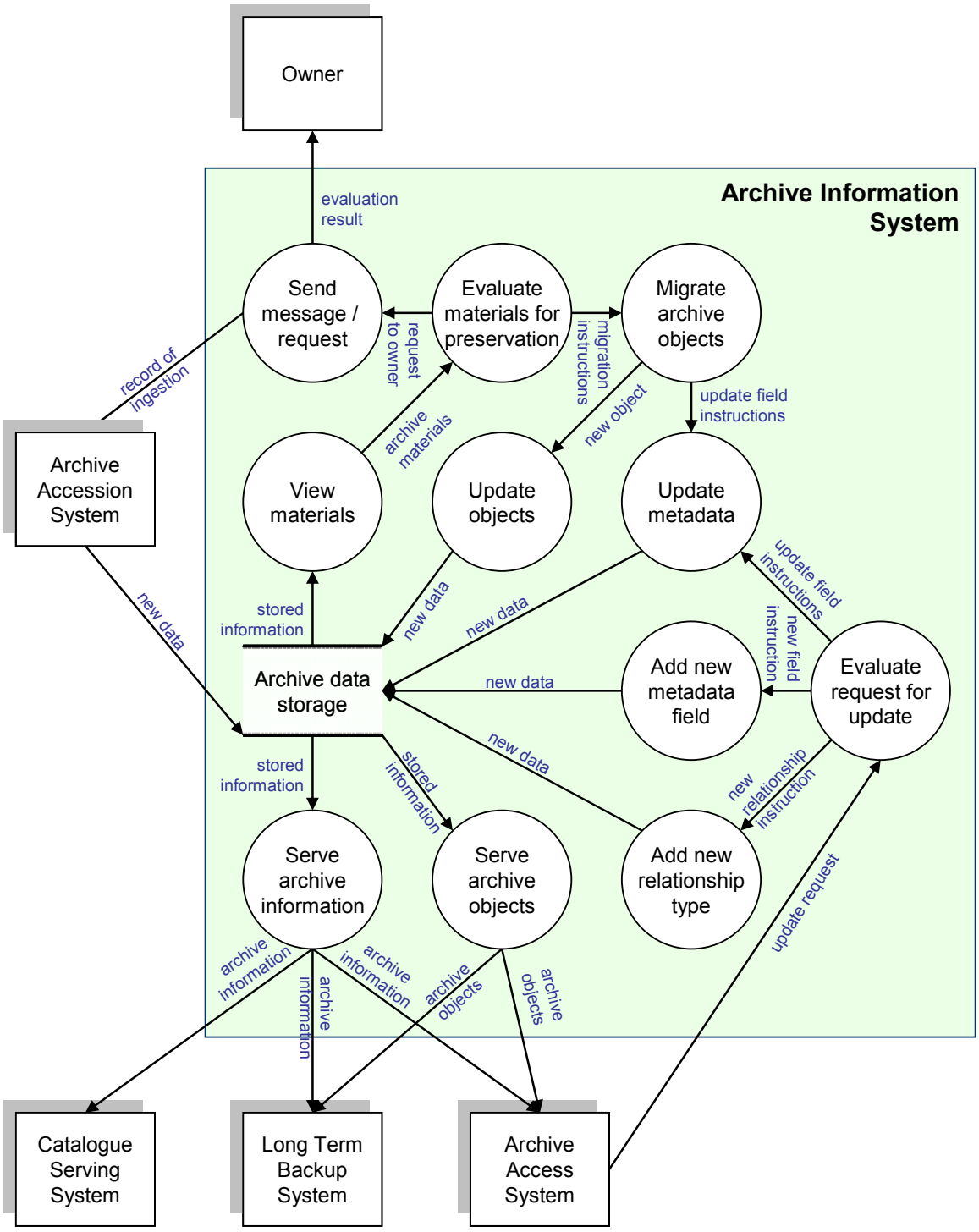


Figure 3.5: Final Data Flow Diagram for archive information system

Physical Model

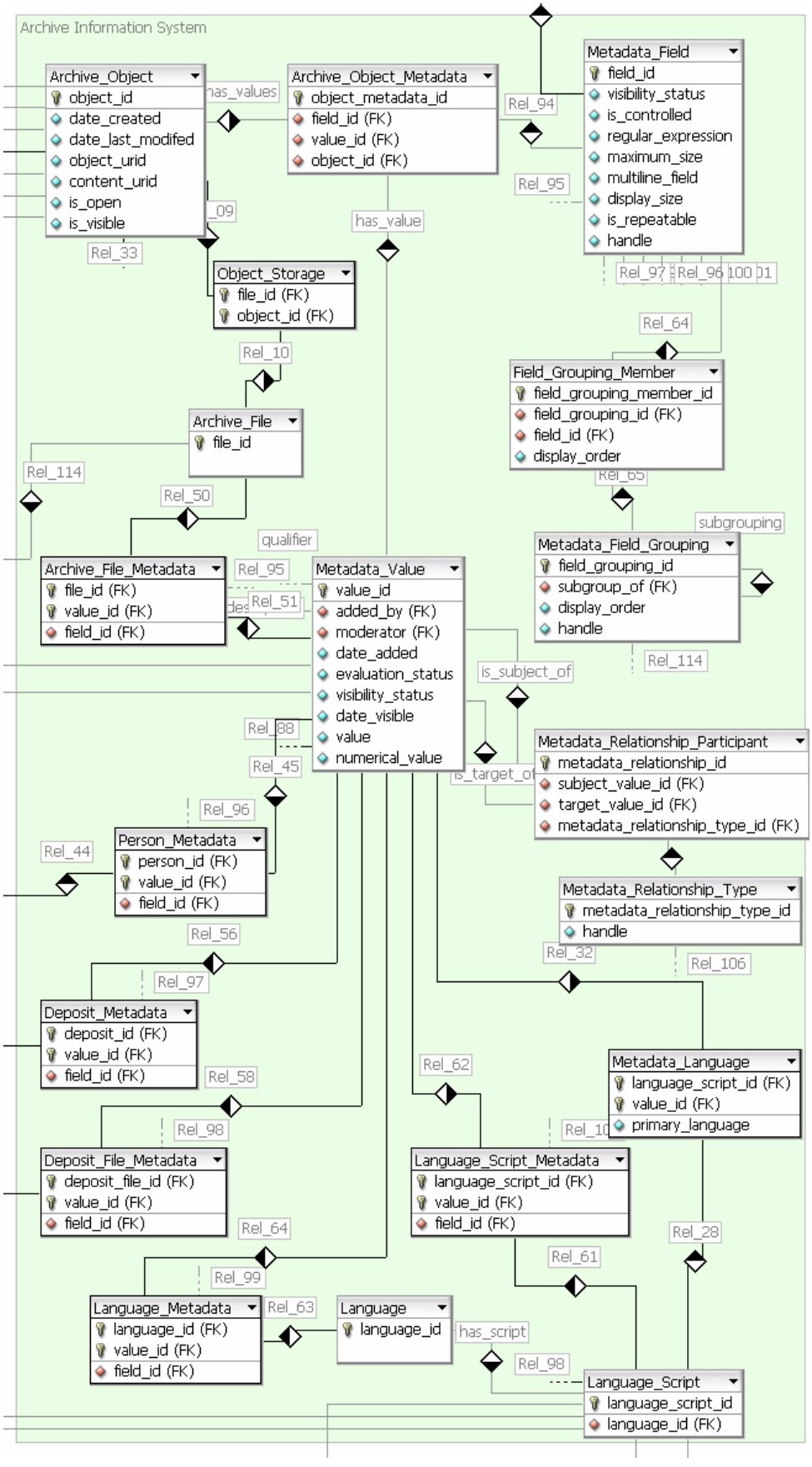


Figure 3.6: Physical model for the archive information system, part 1

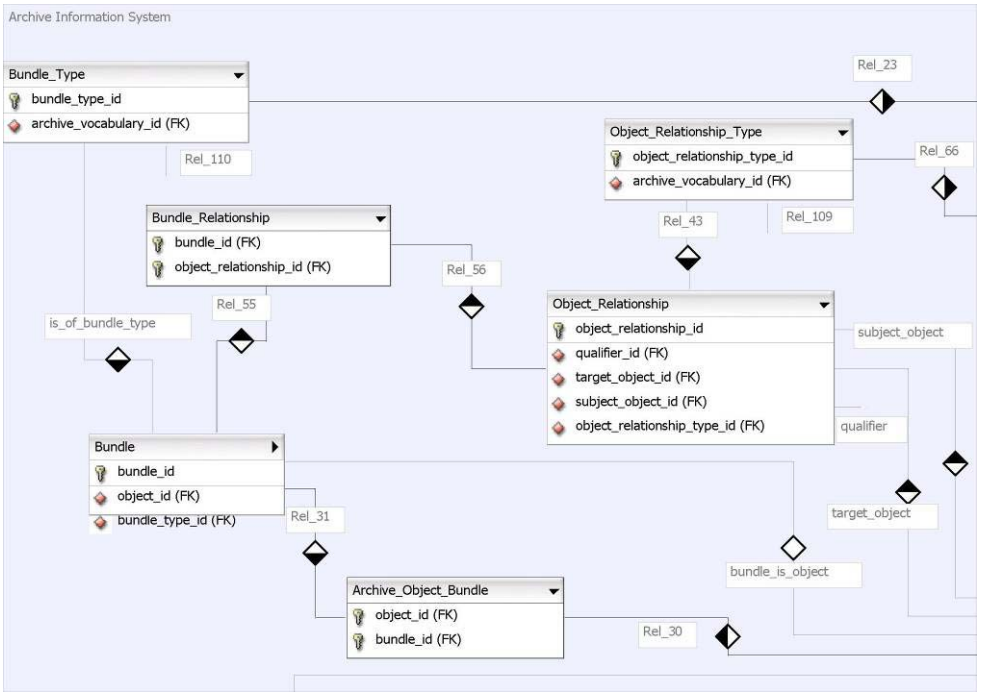


Figure 3.7: Physical model for the archive information system, part 2

Catalogue serving system

Data Flow Diagram

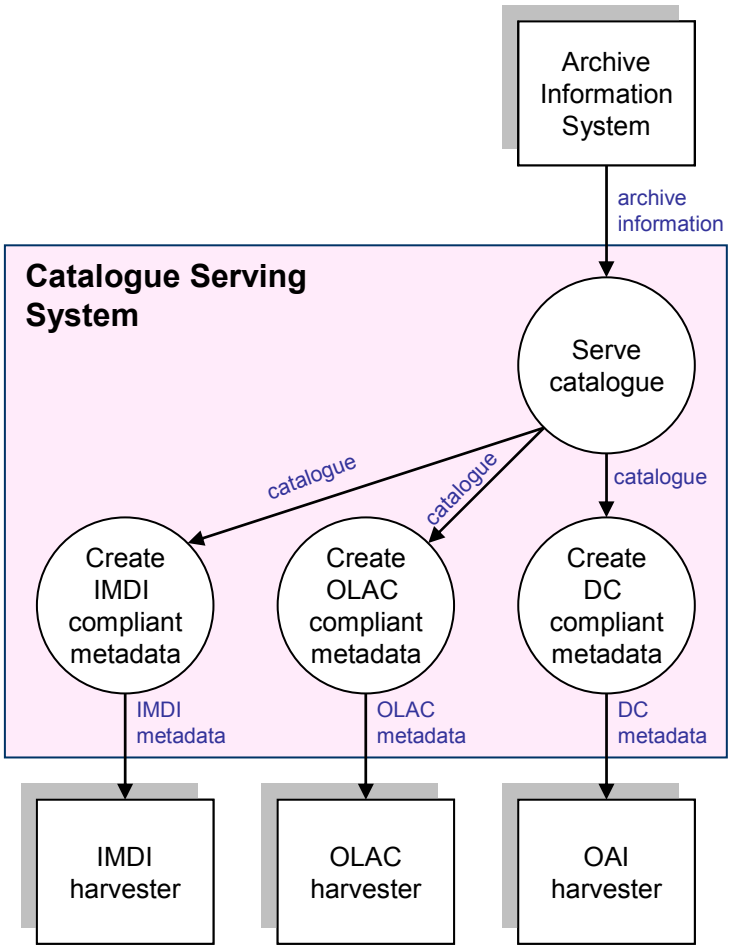


Figure 3.8: Final Data Flow Diagram for catalogue serving system

Physical Model

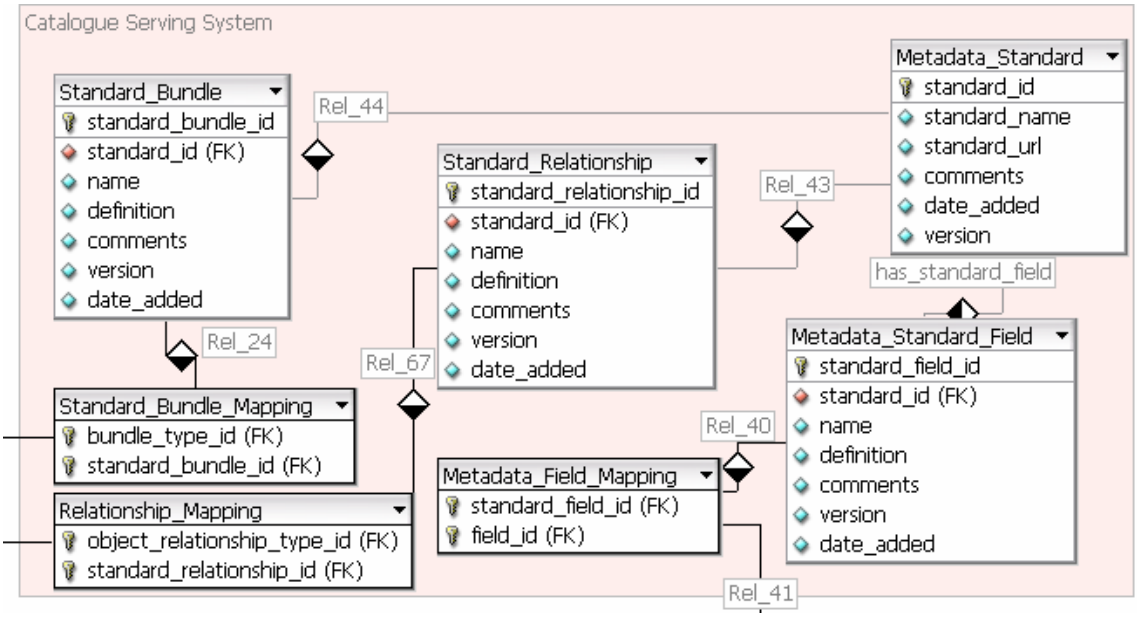


Figure 3.9: Physical model for the catalogue serving system

Archive access system

Data Flow Diagram

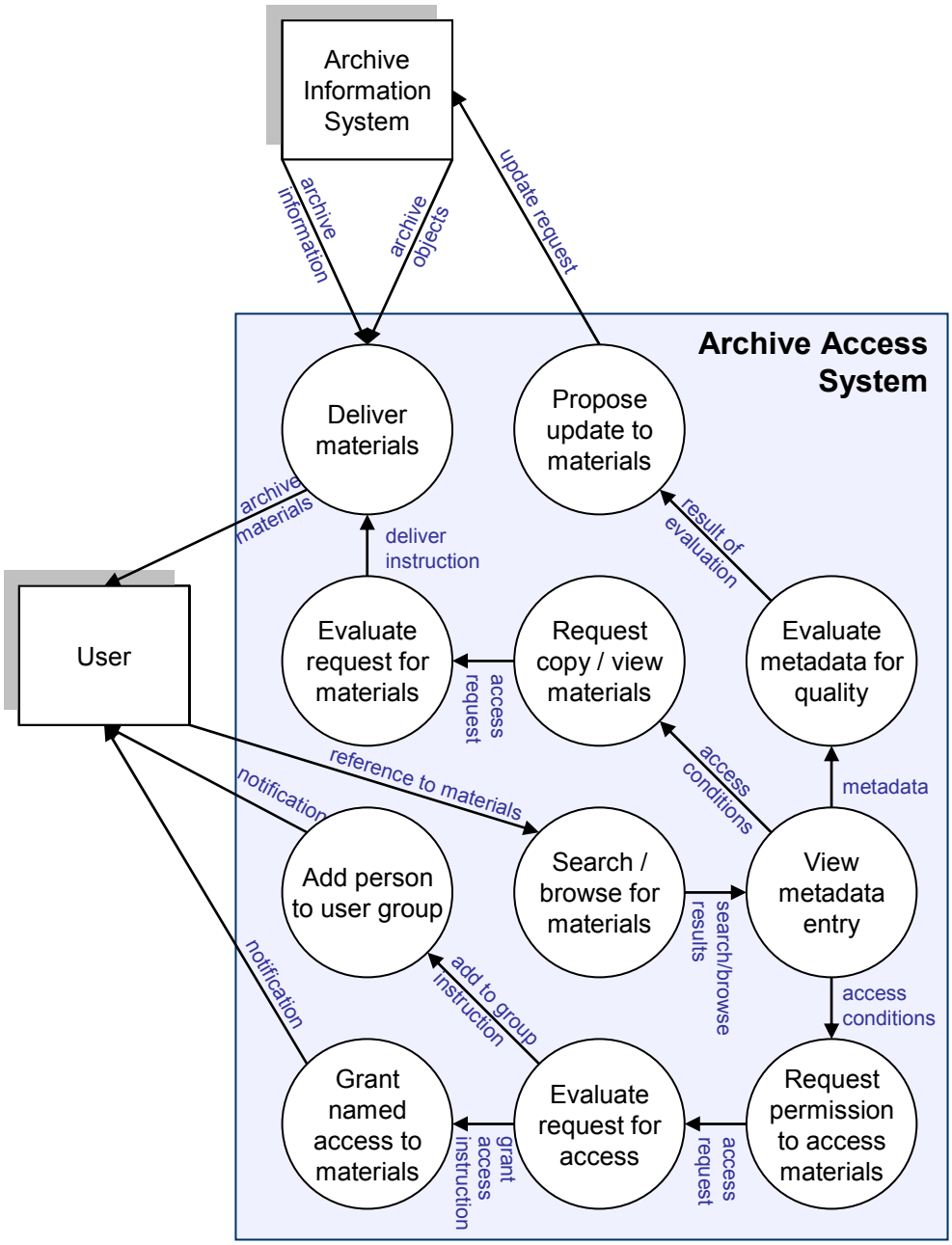


Figure 3.10: Final Data Flow Diagram for the Archive Access System

Physical Model

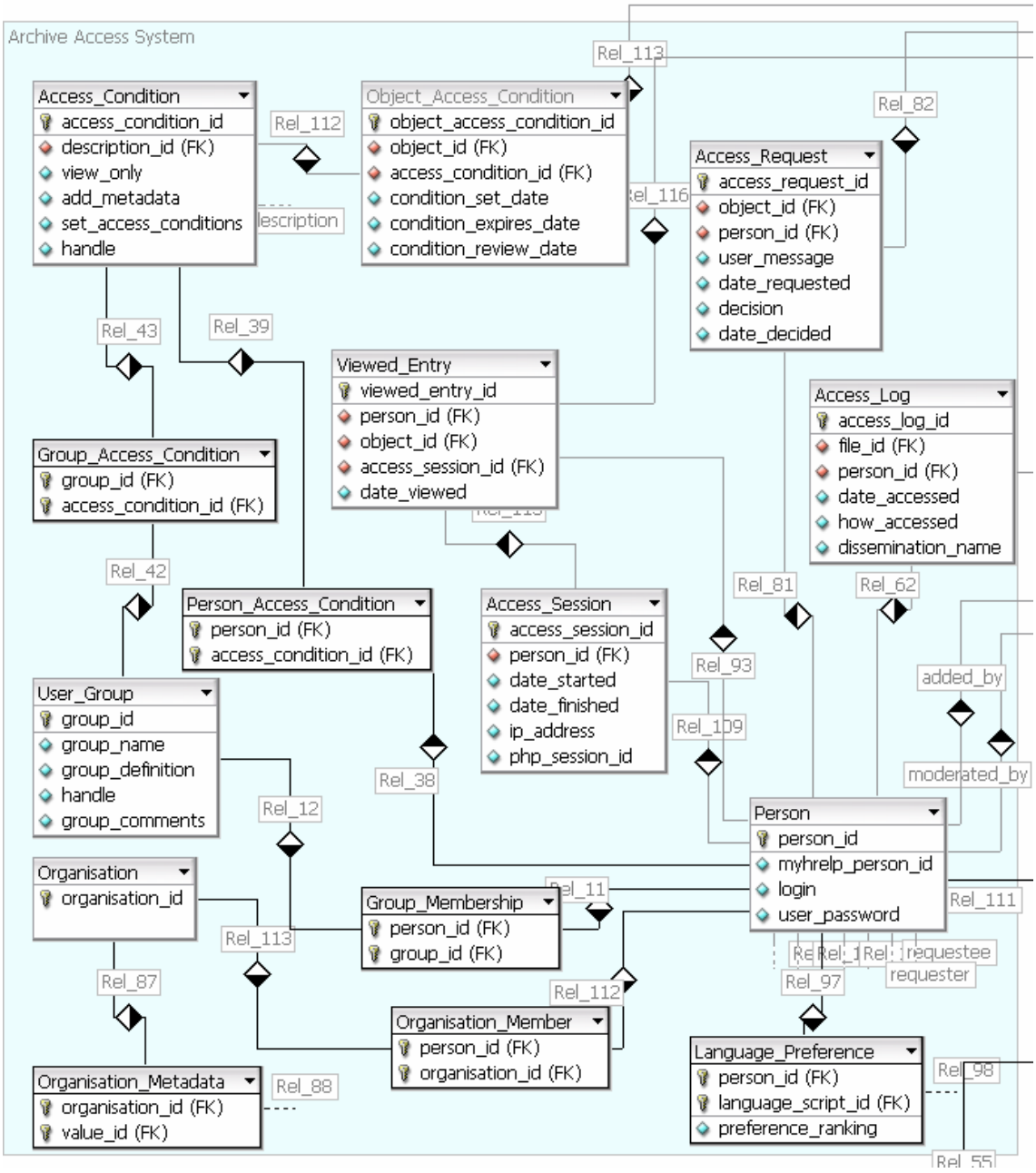


Figure 3.11: Physical model for the archive access system

Management

There were additional tables that in the logical model that related to the ongoing management of the information system than any particular system, mostly containing Data Management Data. They are given below.

Physical Model

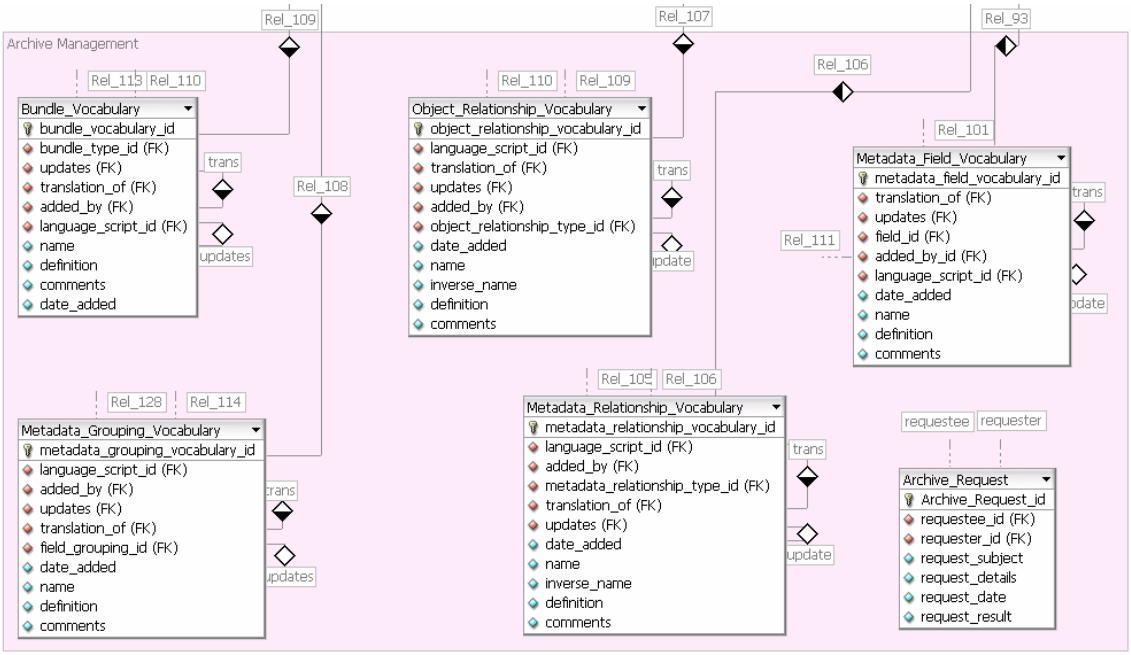


Figure 3.12: Physical model for archive management tasks

Appendix 3: London DAM-LR Meeting Report

DAM-LR Meeting London
Report

This note is a report from the DAM-LR meeting at SOAS in London. The agenda is attached to this document as well as the slides from the coordinator.

Date: 25.8.2006

Place: SOAS, London

Participants:

David Nathan	SOAS London
David Evans	SOAS London
Sven Strömqvist	U Lund
Remco van Veenendaal	INL Leiden
Vincent Wagelaar	INL Leiden
Daan Broeder	MPI Nijmegen
Peter WIttenburg	MPI Nijmegen

1. Formal State of DAM-LR

1.1 Deliverables

The deliverables 4.1, 5.1, 6.1 and 7.1 were discussed and it was agreed to collapse them into one report. It is almost impossible to separate these reports efficiently. The coordinator will inform the PO.

D. Nathan who is in charge of 4.1 and will take the lead to create the joint report informed about the state of contributing and writing. Except a few details and corrections all information was delivered from INL and SOAS. Also U Lund provided most of the information, however, they were asked to rewrite their part to fit with the template that was distributed by Nathan. This will make it easier for the readers to study the document and to compare. It was agreed that the report should be ready before the end of August.

The delay for this report is due to the vacation period (absence of persons) and the unknown status of the final document (one or several reports). All work of the partners that are related with DAM-LR goals are in time (see below).

The next definition report (8.3) was discussed by R. v. Veenendaal. A few topics are missing and were clarified. One of the points discussed and to be integrated was the basis of the federation. Here SOAS already presented an excellent paper, but this needs to be transformed into an agreement with a number of jointly acceptable principles. Since currently there is not enough time to prepare this it was agreed to prepare a first version for the next definition report. Also this report will become ready in August.

1.2 Time Line[23]

The global time line for the distributed solution was discussed again based on updated information and new insights. The following table gives an overview about the steps to be carried out in the coming months based on the deadlines set by the TA. This time line which re-confirms older time line diagrams was accepted by all partners. In the following a few remarks will be made to the main parts.

The first joined metadata domain can already be demonstrated with INL, Lund and MPI. SOAS is currently programming the mapping scheme so that within a few weeks also SOAS metadata will be integrated. This work will be finished in September. Lund and INL will create portals that will also link to metadata from all other institutions so that DAM-LR will provide in total three portals. It is agreed that this work will be finished in October. All partners will add new resources to the integrated domain during the whole course of the project. All partners briefly explained their expectations.

With respect to URIDs still some time is left to install the Handle System although most partners have already started setting the service up. It was agreed that every partner should be ready with installation and tests in November. In December we want to carry out integrated tests, so that in the first months in 07 we can implement a full-fledged resolving system including all partners.

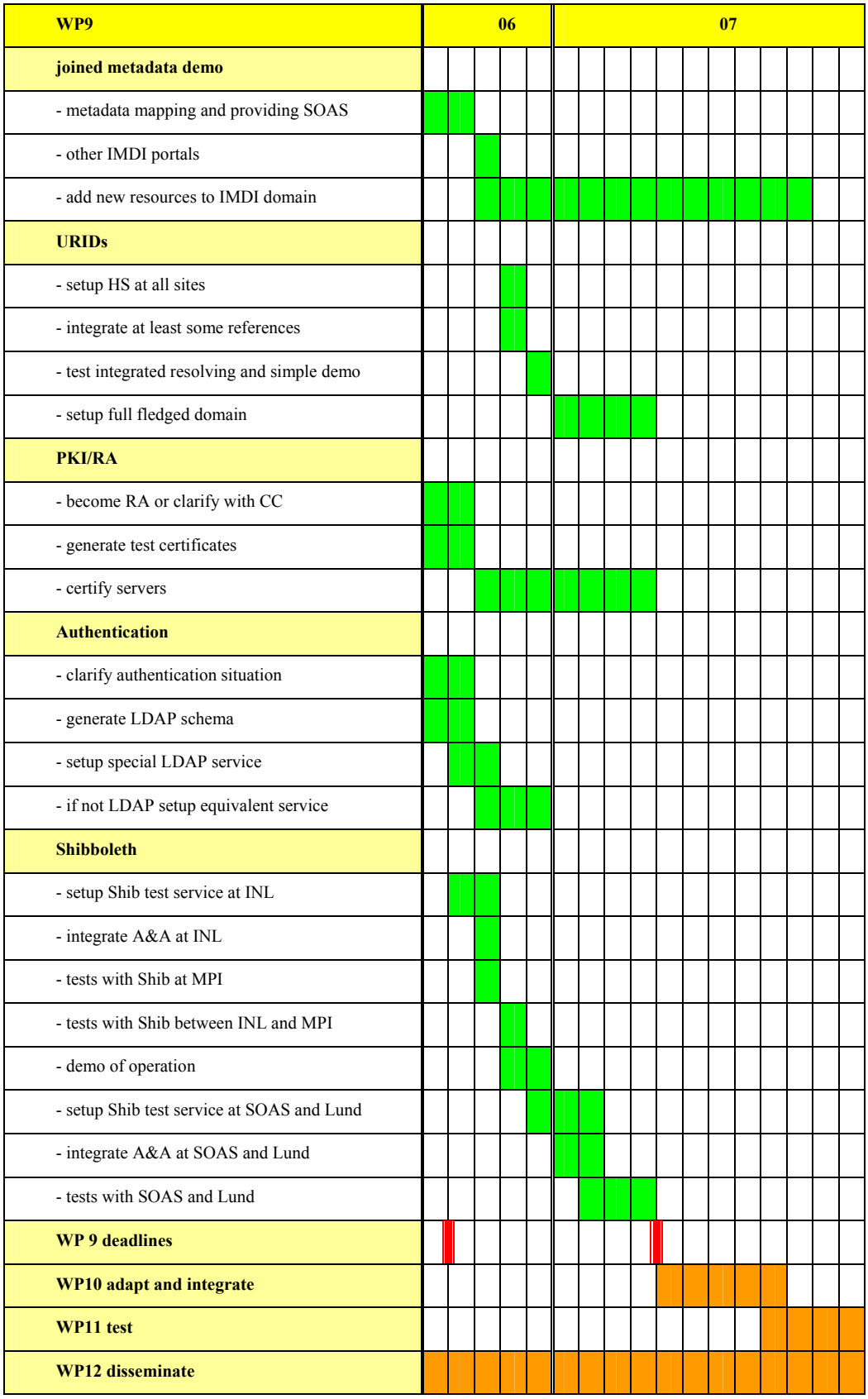
With respect to creating trusted servers with the help of EU accepted certificates all partners have done their work to sort out methods of administration. All matters will be set until September so that all servers being integrated into the DAM-LR domain can be certified dependent on the needs.

Authentication is done differently by the partners since these have to be synchronized with the local computer centers. All partners verified what the local situation is so that they can start implementing a local DAM-LR system supporting all agreed attributes. It is expected that all partners will be ready with their internal setups in October except SOAS. They do not use LDAP and therefore have to make local adaptations to their system. This should be finished at the end of 2007.

Implementing distributed authorization via Shibboleth is a task that will stretch to the deadline in April 07. It was agreed that first MPI and INL will set appropriate mechanisms up where MPI can be used as test case (it is already running and first tests were made successfully). By November it should be possible to a distributed test including INL and MPI so that a demo can be setup and shown in December. In the first months of the new year Lund and SOAS can make use of the existing knowledge and setup their systems, i.e., it will be possible to get a complete domain running at the end of April 07.

This timing diagram was accepted by all partners and will serve as a guideline for the further work in WP 9.

[23] The time line was not yet adapted to a possible demo at the event on 7. March in Brussels. If we will be selected we could adapt it correspondingly.


WP9
06
07
joined metadata demo
- metadata mapping and providing SOAS
- other IMDI portals
- add new resources to IMDI domain
URIDs
- setup HS at all sites
- integrate at least some references
- test integrated resolving and simple demo
- setup full fledged domain
PKI/RA
- become RA or clarify with CC
- generate test certificates
- certify servers
Authentication
- clarify authentication situation
- generate LDAP schema
- setup special LDAP service
- if not LDAP setup equivalent service
Shibboleth
- setup Shib test service at INL
- integrate A&A at INL
- tests with Shib at MPI
- tests with Shib between INL and MPI
- demo of operation
- setup Shib test service at SOAS and Lund
- integrate A&A at SOAS and Lund
- tests with SOAS and Lund
WP 9 deadlines
WP10 adapt and integrate
WP11 test
WP12 disseminate

1.3 Budget Issue

The coordinator reported about the current confusion about budget issues and the difficulties with Form C. It seems that Form C is not filled in correctly, since except for the FCF partner INL there are no entries for the own big investments for which the 10% payment scheme will be applied. There will be a solution. However, the other point is much more serious, since the specifications laid down in the proposal and the TA do not match with the funding scheme. In the TA it is clearly stated that the partners all will make big investments in particular in the first year to set up their own archives. Also in the second and third years there will be local investments, but not that significant as in the first. However, in these two last years it was the intention to ask for EC funds to do the Grid integration, i.e., all the integration work for which we mainly requested EC support will have to be done in the years 06 and 07. If the 10% rule would be applied per year then we would run into big problems and it would not be compliant with the specifications in the TA.

The coordinator was asked to sort out the problems asap with the contact persons at the EC.

1.4 Annual Report and Review

The coordinator informed everyone that we will have to submit an annual report in time, i.e., that he will ask everyone in December to contribute and that everyone has to reserve time for this. Further, he informed about the necessity of making Time Sheets per employee being funded on EC money. A template for these time sheets was distributed.

Every partner was also informed that an official audit will be required this time and that everyone should start discussing now with the administration departments how this is going to be done. SOAS informed us that at SOAS just a complete change took place with respect to budget administration making such things not easy. But the coordinator reconfirmed the importance of the audit aspect.

2. State of the Work

2.1 State at MPI

Daan Broeder reported about the state of the work at the MPI:

- PKI/Certificate issue has been solved
- metadata domain and portal is working and used
- the Handle System is already in operation, however, the metadata infrastructure is set up so that three references can be used in parallel: the traditional way via URLs, the way via internal unique identifiers and the unique identifiers according to the Handle System (to keep the system simple there is a direct mapping between the internal unique IDs and the postfixes of the Handle System)
- the LDAP system is setup such that the Institute-wide ADS system is filtered such that a DAM-LR specific LDAP is filled covering the agreed attributes and allowing to enter external persons that are not members of the institute
- Shibboleth has been setup and the interactions with various components are currently being investigated:
    - Shibboleth identity provider can interact with LDAP and the internal authentication service
    - the JAAS realm of the Tomcat container in which the Shibboleth identity provider is running is used to handle authentication requests
    - Shibboleth resource provider can interact with the Apache HT-Access file

Summarizing one can say that by October all components will have been installed and tested and to a certain extent integrated. The setup will be used to give training courses

etc and it can be used as exemplar installation for the other partners. The MPI is fully on schedule with its work.

2.2 State at INL

The state at INL is as follows:

- The archive has been setup and increasingly more resources will be included. Currently it is the big CGN corpus and the IFA corpus.
- INL has almost finished its work on a metadata portal with a browser running under Tomcat. It was suggested to include metadata references to MPI and Lund; INL did not see problems to do so in the near future.
- With respect to certificates INL will run under the computer center of University of Leiden who is RA.
- A prefix has been requested by CNRI and tests with the Handle System have been carried out already. A postfix syntax has been specified.
- The authentication system architecture is clear, i.e. LDAP will be used.
- With respect to the Shibboleth installation and integration INL would like to participate in a technical meeting talking about specialised matters and use existing experience.

Summarizing one can say that INL is ready for the DAM-LR integration. A software developer will still be around for a couple of months.

2.3 State at Lund University

Lund University has a problem of reserving enough expert capacity since budgets are cut at the university almost continuously. One expert, Marcus Uneson, is occupied with his dissertation work, but Lund will try to ask him to focus on the DAM-LR work in the next period.

- The archive has been setup on a preliminary server and is running. There are ongoing projects to add more data to the archive and make them available via the web.
- A contract for a big department server was just signed, i.e., in a few weeks time the temporary solution will be replaced by a permanent one.
- A metadata portal is available and also Lund will extend the portal such that it will be able to point to all IMDI resources. This extension should not form a problem.
- Lund falls under the RA certification authority of the computer center of Lund University.
- A request for a prefix at CNRI was answered such that it created some confusion. Lund asked for advice and the strategy is very clear: every DAM-LR partner has to request one prefix and has now to pay some money for this. Lund will take it up so that the question can be solved. The postfix syntax is as chosen by MPI.
- The authentication setup is clear: the humanities department will run its own LDAP client with data filtered from the central LDAP system from the computer center. This gives Lund U the freedom to use the DAM-LR attributes and to add new users.
- With respect to Shibboleth the same arguments hold as for INL, i.e., support by an expert would be very welcome.

Summarizing one can say that also Lund University is ready for the DAM-LR integration. But it seems to be necessary to add person power to accomplish the goals.

2.4 State at SOAS

The state at SOAS is as follows:

- The archive has been setup. New data in the order of 0.5 Terabyte is available and will be integrated during the coming months. This then will give a real archive with rich resources about endangered languages.
- SOAS uses a different metadata schema including a number of novel ideas. The schema was presented and it was obvious that it can be finished during the coming week. The intention is to provide mappings and exports to IMDI, OLAC and TEI. A developer is ready to create an appropriate export so that the metadata can be harvested by all portal providers in DAM-LR. Given the limited resources SOAS does not plan to provide a DAM-LR metadata portal itself.
- The certification issues have been clarified with the responsible UK institution (CCLRC) and the appropriate signatures and administrative actions are currently being carried out.
- The Handle System is up and running and some tests have already been done.
- Due to a different setup of the authentication mechanism it is expected that SOAS has to workout exactly how to link to Shibboleth. Here some software development will be necessary that will require deep insight about the Shibboleth identity provider mechanism.
- With respect to Shibboleth the same arguments hold as for INL, i.e., support by an expert would be very welcome.

Summarizing one can say that also SOAS is ready for the DAM-LR integration. However, at the integration side of the distributed authorization the situation will be very different compared to the others.

2.5 General Topic

In the context of work progress it was discussed whether it could be useful for some partners to ask Thomas Soddemann for specialised technical support and knowledge transfer. Peter W will talk to him.

3. Other Aspects

3.1 ESFRI

Peter Wittenburg reported about the ESFRI process which until now turned out to be very successful for CLARIN. Since the German and Hungarian governments already indicated their support for CLARIN it should be sure that CLARIN will get a very high rating and be on the final roadmap. Since the EC will not have so much money it will be likely that CLARIN will go for a preparatory phase. SOAS asked about the scope of CLARIN and Peter Wittenburg stated clearly that the purpose is to stick to Language Resource and Technology providers and that there are major archives with endangered language material for example amongst the current members.

One of the big issues is how CLARIN and DAM-LR are related. The situation is as follows:

- DAM-LR is a Grid project and therefore covers the lower integration levels that are aimed at for a research infrastructure project.
- DAM-LR would count as a bottom-up project and would be rated differently compared to the top-down ESFRI process.
- So it was agreed that Peter should present a follow up DAM-LR II proposal at the next CLARIN meeting in Budapest as the basic layer that will be needed for the CLARIN infrastructure, i.e., it will be fully compliant. DAM-LR II is meant to cover many countries from Europe.
- CLARIN will cover all the other, and without question more difficult to achieve, interoperability layers.

3.2 Significant EC Event

Peter informed the partners about the possibility of participating in a big EC event in Brussels. The partners agreed that we should try to participate and would be willing to speed up certain DAM-LR work to be able to show an integrated domain earlier than planned. Peter will fill in the forms etc.

3.3 Dissemination and Outreach

Peter argued that DAM-LR was very active in its dissemination strategy and that in particular the Live Archives initiative received much interest.

It was discussed what kind of activities should be taken in the next future:

- We will hold a technical meeting (October) about specialised details of the Shibboleth integration. This will be done only for DAM-LR partners.
- In the new year we will offer a 3-day seminar preferably addressing language resource related institutions: one day is reserved for decision makers and managers; two other days will be allocated to discussing architecture and technical details. Lund will prepare a document asap. This should be distributed in wider circles.
- From May 2007 we should offer more training courses for different groups.

A visitor (Tobias Blanke) from AHDS was welcomed at the meeting. In the UK, the ESRC and other organisations are involved in eScience and Grid investigations and exhibitions and DAM-LR was asked to participate. This invitation was kindly accepted and posters will be submitted.

Further, DAM-LR will be represented at the eHumanities workshop of the IEEE eScience conference in Amsterdam in December. This workshop is organized by the CLARIN initiative.

Finally we did not decide on a date for the next meeting, but first wanted to see how the work will progress knowing that different bilateral meetings will be necessary during the next months.

DAM-LR Meeting London
Agenda Proposal

1. Formal State of DAM-LR Peter

9.00 - 10.00

- deliverables and time line
- budget
- annual report
- audit
- review
- time sheets

2. Content Part

10.00 - 14.00

- time line of integration Daan/Peter
- state of work at MPI Daan
- plan for integration tests
    - MD extension and integration plan Peter
    - distributed URID resolving test Daan/Peter
    - distributed shib test Daan/Peter
    - further integration planning Daan/Peter

- state per partner INL/Lund/SOAS

  (focusing on integration aspects)
    - MD schema + mapping
    - PKI/RA
    - LDAP (or other solution for authentication)
    - Handle System - URIDs
    - Shib

- technical support required?
- tech meeting planning Remco

3. Other Aspects

14.00 - 15.00

- state CLARIN - ESFRI Peter
- state DAM-LR II - possible invitation Peter
- dissemination actions Sven/Peter
    - Grid workshop incl. demos
    - Training Events
    - ???

As is obvious to everyone, we have now to derive a clear integration scenario with deadlines. So, please, be prepared for such a discussion - it is crucial now. Please, have a look at the time line proposal from MPI