Deliverable 2.1
Specification of Local Prototype

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: D2.1

Date: 20.6.2005

Content

1. General 2
2. DELAMAN Workshop 3
3. Relevant Initiatives 4
3.1 TEI (Text Encoding Initiative) 4
3.2 ISLE (International Standards for Language Engineering) 4
3.3 ISLE/IMDI (ISLE Metadata Initiative) 4
3.4 OAI (Open Archives Initiative) 4
3.5 Dublin Core 4
3.6 OLAC (Open Language Archives Community) 5
3.7 ISO TC37/SC4 (Language Resource Management) 5
3.8 W3C (World Wide Web Consortium) 5
3.9 GGF (Global Grid Forum) 5
3.10 Grid Inspired 5
4. Specifications for the Local Prototype 6
4.1 Metadata Layer 6
4.2 Physical Layer 6
4.3 Access Management 6
4.4 Ingesting Resources 7
4.5 Archive Management 7
5. State of the Local Prototype 7
Appendix A: Workshop on Access Management 8

1. General

The framework within which DAM-LR was defined can be characterized by a number of facts and trends:

- It is much easier to create a large collection of multimedia language resources and to maintain them on computers.
- To enable re-usage of language resources they will be stored in central archives that have the capability of maintaining them for many years, organizing them according to clear principles and giving access to them via the Internet.
- The size of language resource collections is growing rapidly. This is true for textual corpora, but in particular for collections that cover sound signals such as the Dutch Spoken Corpus (CGN), or especially these that contain also video recordings to document multimodal communication.
- Individual researchers or projects will increasingly input their resources into central repositories. In particular where recordings are unique, as is the case in the documentation of endangered languages that will become extinct in a few years, it will become a duty to store the data at central places. They have to solve the long-term preservation task.
- The creation of linguistically motivated metadata is more often accepted to make resources available to the interested community, to document non-linguistic parameters, to organize and bundle resources and to manage them.
- Increased visibility of resources via publicly available metadata will foster the user's desire to be able to access and utilize them for different purposes – be it linguistically, technologically, educationally or other reasons.

Envisioning these trends' interoperability is a big concern and one of the major challenges for the future. This is fully supported by all ideas in the Semantic Web. Interoperability can be assigned to different levels:

- Semantic interoperability concerns unifying and relating the vocabularies that are used to characterize domain specific phenomena. Due to differences between linguistic theories and languages, all projects use different tags and encodings to characterize linguistic phenomena.
- Structural interoperability relates to unifying and relating the structures/formats in which language resources are packaged.
- Technical encoding interoperability is about unifying and relating the encodings that are used to represent characters, sound streams, images and video streams.
- Organizational interoperability is about making it possible for users to access different repositories of language resources as if they were one repository.
- Expecting that web services will be offered increasingly often, functional interoperability will become an important issue.

DAM-LR is located at the level of organizational interoperability, i.e. it creates

- one integrated metadata domain based on the IMDI standard where the users can browse and search to discover useful resources
- one domain of unique identifiers for all resources covered
- one domain of user identity so that they can operate in the archives with a single sign-on
- one point of authorization-granting so that for these user identities access authorization to the language resources is defined, independent of where the resources are offered.

2. DELAMAN Workshop

In November 2004 the MPI organized an international workshop[1] entitled "Access Management for Distributed Archives" under the umbrella of DELAMAN[2] which is a world-wide network of archives with endangered languages and music material. DAM-LR can be seen as a project that intends to implement the part of the DELAMAN goals which have to do with creating an interoperable archive domain. Therefore, the workshop was an excellent preparation of the work that has to be carried out in DAM-LR. All partners were represented and took active part in the discussions.

In this report we only want to briefly summarize the main results of the workshop, for more details we refer to the workshop report (see appendix A) and the papers presented. About 30 experts (Researchers, Archivists, Technologists) from leading institutions were invited to discuss user and archivists' requirements, available and emerging technologies. Amongst the technologists were experts from the leading initiatives to get an excellent overview of, and deep insight into the potential of software components. The major results are briefly listed:

- The DELAMAN archives want to federate their collections, i.e. they want to
  - exchange the data to ensure long-term preservation
  - maintain the access policies for all copies of the resources, and leave the control of them at the originating archive
  - intergrate and share the user base with one accepted identity per user
  - make it possible for interested communities to have a single sign-on when using material from the DELAMAN archives
  - create a domain where users can create their own virtual metadata domain based on resources from different archives

- The DELAMAN archives want to build on existing tools and frameworks were possible and there is excellent software at various levels such as
  - Web services as an underlying basis for software architectures
  - A-Select for layered authentication
  - Shibboleth for distributed authorization
  - the Handle System for resolving unique identifiers to physical paths
  - Grid components for various aspects, such as monitoring access to databases
  - Storage Resource Broker as a comprehensive solution for federated collections

[1] The workshop was funded by the VolkswagenFoundation – www.mpi.nl/delaman
[2] Digital Endangered Languages and Music Archive Network – www.delaman.org

- The DELAMAN archives want to start an evaluation and decision process about the following topics:
    - What do the different solutions do and what do they not do?
    - Which products can work together?
    - Which layer to choose, i.e. start from SRB and carry out adaptations or start basically from scratch with for example WS components?
    - Which type of basic infrastructures have to be setup, such as a PKI system?
    - How can we integrate the different metadata concepts?
    - etc

The partners in DAM-LR, and in particular the local prototype, play an important role in the DELAMAN process. It will be used to study a number of technology components and to analyze and test their potential.

3. Relevant Initiatives

Here we would like to briefly mention some of the work that is currently carried out in the initiatives that are relevant to DAM-LR.

3.1 TEI (Text Encoding Initiative)

The focus of TEI is mainly on making suggestions for structuring linguistic resources and for encoding all types of linguistic phenomena – so it addresses interoperability layers that are not in the focus of DAM-LR. However, TEI also makes suggestions for so-called header information which are metadata categories to characterize linguistic resources. These are relevant for the organizational interoperability layer, and some resources covered under DAM-LR such as the Dutch Spoken Corpus have made use of them. The IMDI metadata standard can incorporate TEI header categories and special profiles can be created.

3.2 ISLE (International Standards for Language Engineering)

The ISLE was a follow up project to the EAGLES project and had as a goal to define linguistic concepts used in the area of language engineering, and come up with a number of standards such as in the area of multilingual lexica. For DAM-LR only the metadata work is of direct relevance (see below).

3.3 ISLE/IMDI (ISLE Metadata Initiative)

One topic of the ISLE project was the definition of a metadata standard and an appropriate infrastructure supporting the standard. The work resulted in a complete IMDI infrastructure that allows browsing and searching. Therefore it can be used for resources discovery based on a domain-specific set as well as for resource management. For DAM-LR it is one of the pillars.

3.4 OAI (Open Archives Initiative)

The focus of OAI is to discuss architectures, characteristics and components of open archives and work out proposals. Most archives are compliant to the architecture of open archives, since it is a very abstract model. The local prototype is fully compliant with what is proposed. Of great importance is the OAI Protocol for Metadata Harvesting (OAI-PMH) which is used as a lightweight protocol to exchange metadata. One of the criteria for PMH compliance is to deliver Dublin Core metadata records as a default. The IMDI infrastructure is supporting the OAI-PMH protocol and provides mappings to the DCMES semantics.

3.5 Dublin Core

The Dublin Core initiative worked out a metadata set (DCMES) for web resources that includes 15 broadly and vaguely defined elements. The goal of DC was to cover the most important semantic aspects that may be sufficient for the discovery of any web resource. Therefore, DCMES will be relevant for general discovery services. This is the reason why all DAM-LR archives will act as metadata providers for any interested service provider and deliver DCMES records.

3.6 OLAC (Open Language Archives Community)

OLAC is an initiative that extended and specified the DCMES for the domain of language resources to allow more specific queries. Two elements were added that are crucial for the discovery of language resources and some refinements were made while preserving the DCMES packaging. Language resource archives should provide OLAC/DCMES compliant records to offer a large searchable domain for language resources.

3.7 ISO TC37/SC4 (Language Resource Management)

The goal of the recently formed ISO TC37/SC4 is to define standards for the proper management of language resources which can be about structuring resources, encoding linguistic phenomena or characterizing language resources. Most aspects are addressed that tackle other interoperability layers than are dealt with within DAM-LR. It is important that the IMDI metadata set and the OLAC/DC semantics will be integrated into the metadata part of the ISO Data Category Registry (DCR) so that other initiatives will be able to make use of the definitions in a flexible way.

3.8 W3C (World Wide Web Consortium)

W3C worked out many very useful and widely used standards such as XML which are essential for the representation of archival contents of all sorts. It is evident that DAM-LR archives will make extensive use of XML and schemas to structure textual documents. W3C also defined the Web Services (WS) which is another strong pillar for establishing the Semantic Web. It defines a general web-based way to describe services and to exchange structured messages. Another very important result of the W3C efforts are the Resource Description Framework, its schema variant and the OWL developments. Currently, descriptive metadata is defined in XML format, the ISO DCR is also represented in XML. However, for reasoning about web-based metadata resources these technologies will become more relevant. DAM-LR will apply the Web Services technology where necessary and possible. The RDF technology does not seem to be relevant at this moment for DAM-LR.

3.9 GGF (Global Grid Forum)

GGF is a broad initiative that creates architectures and develops software for connecting computers and computing services at different levels. It originated from the idea of Grid Computing where the intention was to connect distributed computer centers to share big computer intensive jobs. It emerged in speaking about distributed computing in all respects. Web Services were extended to support state, middleware and workflow components which are being developed and architectures designed that are relevant for the DAM-LR topics. However, the recent DELAMAN workshop also showed that GGF is fairly scattered and that the software components are not yet ready and stable. Nevertheless, DAM-LR has to look into the GGF deliveries.

3.10 Other relevant Initiatives

Many software development initiatives emerged in the realm of the Grid, Semantic Web and Digital Libraries initiatives. Here we can mention only a few that certainly have to be analyzed with respect to their usefulness for DAM-LR:

- **Handle System**
  The Handle System was designed by CSRI to meet the increasing needs for efficiently handling unique resource identifiers in a domain of distributed archives. It is mature software now that can be integrated into various frameworks.
- **A-Select**
  A-Select was developed by SURFNet as a professional authentication system with multiple levels of security. It is mature software and can also be integrated into various frameworks.
- **Shibboleth**
  Shibboleth was developed within the I2 initiative to meet the requirements of authorization in a domain of distributed resources and distributed access granting authorities. It is mature software and can be integrated into various frameworks.
- **SRB (Storage Request Broker)**

At the Supercomputing Centre in San Diego the SRB software was developed in the last few years to allow for the creation of federated archives. It has many features in it that are interesting for DAM-LR and it is robust software.

- **D-Space**
  D-Space is a software for federated digital libraries developed by a large consortium under the leadership of MIT. Also this package is interesting for DAM-LR.
- **Fedora**
  Fedora is another development for federated digital libraries developed ...

There are many other components and systems around and not all of them will be investigated by DAM-LR and not all of them were considered when specifying the local prototype. What has to be checked is how much work can be put into the analysis of existing solutions.

4. Specifications for the Local Prototype

The local prototype to be developed at the MPI for Psycholinguistics has to adhere to the following specifications:

4.1 Metadata Layer

- All archive resources must be described by IMDI metadata.
- All IMDI descriptions must be open and accessible as XML files.
- Structured and unstructured metadata searches within the IMDI domain have to be possible.
- Searches with the help of search engines such as Google have to be supported.
- The IMDI resources have to be offered as OLAC/DC records according to the OAI PMH protocol to make them searchable via OLAC services providers.
- The IMDI files must be linkable into a linked domain that supports an organization of resources into logical bundles, browsing and management.
- It must be simple to register and integrate new IMDI-based repositories into the IMDI domain.
- It must be easy to setup an IMDI portal.
- If authorized it must be easy for users to access resources via the metadata descriptions.

4.2 Physical Layer

- The physical storage must be transparent to the user, i.e. the user should not have to deal with servers, disks etc.
- The physical location of the resources should be easily modifiable without causing problems for the users.
- The organization of the archive should allow for the copying of whole and sub-parts of the archive to new archives to support long-term preservation and redundant access paths.
- Each archived resource has to be identified by a unique resource identifier (URID). The metadata descriptions have to contain URIDs to refer to the resources.
- The storage concept must be such that several copies of all resources can be generated automatically and location resolving can be carried out.
- The stored resources have to be in archivable formats and directly accessible for authorized people.

4.3 Access Management

- The access management system must support the definition of policies (declaration of code of conducts, usage, processes etc) and rights.
- The access management system must support the specification of usages and temporary tickets associated with these usages.
- The access management system must support efficient electronic operations via web interfaces and a delegation mechanism to allow resource owners to define access policies from remote sites.
- Access policy specifications must be based on the metadata layer, i.e. the physical layer is transparent to the definer and the specifications are independent of the physical location of the resources.

- It must be possible to specify domains of authority in the metadata layer.
- The delegation of rights must be possible.

4.4 Ingesting Resources

- The possibility to integrate new resources or update existing resources into the language resource archive has to be controlled by an upload system that ensures that its coherence and consistency is guaranteed.
- The user has to be provided with a workspace mechanism which allows him/her to arrange the data and test its compliance until it is ready for integration.
- The upload system has to be equipped with a configurable list of permitted file types and where possible with format checkers. In particular for complex resource types, dependent types must be indicated, some being required such as a schema.
- The upload must support the definition and integration of an upload node in the existing archive, archive structure, metadata descriptions and resources. It must support the proper linking of these elements.
- Versioning must be done in the case of integrating new versions.

4.5 Archive Management

An archive management shell has to be available to carry out typical management activities such as:

- copying and moving data while retaining the correctness of the archive's organizational links
- checking the consistency of all links in the archive and modifying them where necessary
- checking the format and technical encoding correctness of all resources where possible
- automatically generating additional resource types for presentation purposes such as MP3, MPEG4, etc
- creating different types of statistics
- the possibility of removing sub-parts of the archive which is the most dangerous operation and which therefore has to be guided

5. State of the Local Prototype

This chapter was added to give an indication of the typical size of a modern language resource archive and the dimensions of the problems. Since the archive is continuously changing and increasing in size we can only give rough numbers.

The archive at the MPI for Psycholinguistics contains

- about 43.000 annotated media files (bundles of video and sound files)
- about 100.000 annotation files
- a large number of notes and additional types such as lexica, grammars, field-notes, etc
- about 190.000 archival objects
- about 30 authority domains

Appendix A: Workshop on Access Management

International Expert Meeting on Access Management for Distributed Archives

Report

Peter Wittenburg
5.2.2005

On November 29/30. 2004 a group of distinguished linguists, language archivists and technologists met at a workshop about Access Management for Distributed Archives in the realm of the DELAMAN network (Digital Endangered Languages and Music Archives Network). This report gives a summary of the talks and discussions. They will guide the future work within DELAMAN and will have consequences for the DAM-LR (Distributed Access Management for Language Resources) project funded by the EC starting officially at 1.1.2005.

After an introduction to the goals and a contribution about archiving[3] first the researchers needs and ethical and legal considerations were presented. In the second slot language archivists presented their current architectures. Afterwards these requirements were summarized and discussed, before an overview was given about relevant technological aspects. After this preparation a number of technologies and products were presented by technologists. Before some final conclusions were drawn, rounds of discussions reflected on the different presentations and on whether the solutions fit with the linguists' and archivists' expectations.

Content

Workshop Goals 9
Workshop Results 9
Archiving 11
Researcher Requirements 12
Ethical and Legal Aspects 12
Archivist Requirements 13
Technologies and Products 15
References/Glossary 19
Acknowledgements 20

MPI's managing director, Wolfgang Klein, welcomed all workshop participants and explained the natural interest of the Max-Planck-Institute in the topic addressed by the workshop. MPI researchers are also involved in creating digital corpora to support their research. He pointed out that probably less than 5% of the world's languages are well-studied, and that the most urgent problem of modern linguistics is the lack of reliable empirical evidence on the vast majority of languages. The issues discussed at this workshop and work following from it should be seen as a major contribution to the solution of this problem.

We were very happy that Wilhelm Krull, Secretary General of the VolkswagenFoundation (VWS), participated in the workshop, since VWS is one of the major funding institutions currently involved in supporting the documentation and archiving of endangered languages. In his welcome talk he presented the activities of his foundation, the reasons why endangered languages were chosen as a topic of support and described his expectations. Efforts should be taken to solve the long-term preservation aspects on the one hand, on the other hand the material has to be accessible to all interested communities.

[3] The exact meaning of the concept "archiving" in the area of language resource preservation was subject of discussion (see below).

Workshop Goals

The background of the DELAMAN discussions is the emergence of increasingly larger digital archives in the area of endangered languages and music that store data that will not be available at a later moment in time since these languages will become extinct in a few years. This means that we have a severe task to solve the long-term preservation of the material. Further, we know that material about certain languages or language families are housed in different archives. Therefore, we understand the wish of the researchers and language communities to see an integrated domain of language resources. The Internet allows us to tackle both aspects. However, problems of various sorts have to be tackled and solved.

In his opening talk Peter Wittenburg briefly sketched the very short history of speaking about computer-based digital collections in the area of language resources and its requirements such as metadata, generic annotation formats, lexicon structures, character representations and others. It was the LREC conference in Athens that had the first workshop on these issues and that was the moment that several excellent initiatives were started. Only four years later we can observe emerging collections stored in language resource archives in the area of endangered languages and music such as AILLA, DOBES, ELAR, LACITO, MPI, PARADISEC, and also in the area of human language technology. We also see that traditional archives such as the national sound archives started to house digital material as well. In this short time we could come to widely established agreements about open standards for archiving such as UNICODE, XML, WAV, MPEG2, JPEG, TIFF and others although it is understood that still some areas, such as the representation of images and video, are subject to dynamic changes. With respect to metadata descriptions of language resources two main proposals with different foci are widely used (IMDI, OLAC) and these interact with each other with the help of gateways. Still there are a huge amount of digital language resources that are still not visible and catalogued.

Wittenburg presented his view that these well-organized and highly accessible archives become increasingly important and that not only linguists but also others are starting to accept the role of the central repositories as a place where results are stored reliably. He also described the necessity that language archives are not static and must offer advanced access methods, since researches for example want to revise secondary data such as annotations or lexica due to new insights and others such as members of the language community want to comment on the content. He briefly sketched the ideas that were discussed in the DELAMAN network and that are part of the coming DAM-LR project. These influenced the goals of the workshop:

- We needed to get a deeper understanding of what potential users expect from the emerging DELAMAN archives.
- We needed to get a good view about the ethical and legal aspects that are involved when implementing the DELAMAN goals and must not be overlooked.
- We needed to understand how the existing archives are organized and what has to be done to realize the DELAMAN goals.
- We needed to get a detailed view about current technology trends and about the state of the software tools and frameworks that are advocated, since they may form the basis for implementing the DELAMAN goals.

Workshop Results

The primary result of the workshop is that it managed to draw an excellent overall picture about the goals we want to achieve and the aspects that have to be solved along that way. There are user-related, legal and ethical and technological aspects. The workshop also resulted in a number of concrete steps that have to be tackled in the future. In general the organizers got excellent feedback which encourages the DELAMAN people to move ahead with their plans.

Results

The goals we want to achieve within DELAMAN stabilized and there seems to be a wide agreement between the archives. This is mirrored in the results of the workshop:

- We want to federate our collections, i.e. we want to
    - exchange the data to ensure long-term preservation
    - maintain the access policies for all copies of the resources and leave the control about them at the originating archive
    - an integrated and shared user base with one accepted identity per user

    - make it possible for interested communities to have a single sign-on when using material from the DELAMAN archives
    - create a domain where users can create their own virtual metadata domain based on resources from different archives

- We want to build on existing tools and frameworks were possible and there is excellent software at various levels such as for example
    - Web services as an underlying basis for software architectures
    - A-Select for layered authentication
    - Shibboleth for distributed authorization
    - the Handle System for resolving unique identifiers to physical paths
    - Grid components for various aspects, such as monitoring access to databases
    - Storage Resource Broker as a comprehensive solution for federated collections

- We have to start an evaluation and decision process about the following topics:
    - What do the different solutions do and what not?
    - Which products can work together?
    - Which layer to choose, i.e. start from SRB and carry out adaptations or start basically from scratch with for example WS components?
    - Which type of basic infrastructures have to be setup such as a PKI system?
    - How can we integrate the different metadata concepts?
    - etc

After the workshop we shared the feeling that there are good options and that it should not take too much time to find answers to these questions. This fits with Egon Verharen's remark, that components are available and that the rest is a matter of decision and integration.

With respect to a number of topics such as a common ethical ground we have to involve our donators and the language communities before we can make final agreements.

It is obvious that we should establish good contacts with communities such as organized in IASA, in particular, since it is not clear what the exact division of labor will be in the future. Currently, in our domain "archiving" and "access" are closely related and can't be separated. But we cannot make statements whether this will still be true in a few years.

A number of action points were made and responsibilities were defined (the persons mentioned are in charge, but will probably form small working groups):

- **Gary Holton** will check the relevant ethical documents and write a document that has to be discussed broadly and then can be used as a common ethical ground.
- **Heidi Johnson** will check the legal aspects and write a memorandum that has to be discussed and then can be used as common legal ground.
- **Peter Wittenburg** will check the aspects of access management and write a document that has to be discussed broadly and then can drive the evaluation and implementation work. Aspects are the integrated user domain, single sign-on, access policies, access rights levels, etc.
- **David Nathan** will work out an improved version of usage scenarios that will tell us in more detail which services the DELAMAN archives have to give.
- **Linda Barwick** will work out a DELAMAN constitution and will also work out a document that describes the type of extra work that is involved when the archives will integrate into a joint infrastructure that has to be maintained.
- The **DAM-LR** folks see DAM-LR as a test-bed for DELAMAN intentions and will report about the decisions and experiences. Peter Wittenburg will take care of this exchange of information. It was seen as useful when also in other countries funds would become available so that DAM-LR could more actively collaborate with the non-EU colleagues in DELAMAN.
- In **DAM-LR** we will start looking[4] at things such as

[4] If there is no special request we will not look into the Handle System issue again. It was tried out at MPI and works perfectly. The only question is whether it can be easily integrated with other components that will be chosen.

- common PKI and certification infrastructure
- authentication and authorization frameworks
- federation options

With the exception of the last two points it was suggested to have first notes ready in March. Due to the slightly delayed appearance of this report, we should shift the deadline to April. Since for example all DOBES teams will meet in May it would be excellent to have first ideas ready then, since we could present and discuss draft documents to get feedback.

The next DELAMAN meeting will be in November 2005 in Austin and at that date we should have elaborated documents. At that date there should also be some evaluation reports about the last two points mentioned in the action point list. It was also suggested recently to bundle these documents into a book to document the process and make it transparent to all parties.

Archiving

Dietrich Schüller was invited to present his ideas about “Audiovisual Archiving: Visions, Challenges, Strategies” being one of the key persons in the corresponding IASA (International Association of Sound Archives) discussions. Audio(visual) archives were already founded from 1899, i.e. there is already much experience. With the advent of small battery-driven audio and video recorders the amount of audio and video carriers exploded within shortest time frames and is now expected to be around 150 Million hours. Based on their experience Schüller made a distinction between archives and dedicated data bases. According to him archives have to take care of long-term preservation, while databases focus on access and presentation.
It is obvious that most of the recorded material exists outside of archives and even linguistic research institutions. It will become a huge task to spot a fragment of these recordings and safeguard them, since most of them are endangered due to different types of degradation, format obsolescence and a lack of suitable equipment. Lots of a/v formats can hardly be handled anymore and new storage media such as recordable CDs and DVDs have only a very short life-time. Increasing storage densities also lead to a decrease in data persistence. Therefore, around 1990 sound archivists recommended a change of the preservation paradigm: The long-term preservation of original documents is hopeless! Only the lossless copying of digital documents to new media will help. Digital Mass Storage Systems are the primary preservation solutions for broadcast and national, followed now also by research archives.
He briefly touched a few principles for digitization: 48 kHz/24 bit is seen as the minimal resolution for audio, 96 kHz/24 bit is the standard for cultural heritage objects and 192/24 bit is upcoming for digitizing mechanical originals. The basic principle is: The worse the signal, the higher the resolution. Under the prevailing low budgets, the transfer of all available recordings to digital storage alone will take decades and will cost a huge amount of money. For many old recordings it will become an art to find, or reconstruct useful equipment and to adjust it properly.
Further, he proposed a number of strategic aims: (1) A considerable amount of funds have to be provided for digitization and maintenance of digital collections; (2) All material should be ingested into adequately equipped and experienced archives since individuals and many institutions that house about 80% of the material in the area of ethno-linguistics lack basic skills and capabilities; (3) There should be a division of labor between archives in the narrower sense and providers of content related databases. The aim of archiving is giving access to information: however, access must be based on the physical existence of the material. “Ingest”, the professional conversion of contents from original carriers of any kind of archival digital formats, and thereafter maintaining their physical existence in the long term is not a trivial task as history has shown. Finally he recommend a close collaboration between DELAMAN members and the IASA committees.

Discussion
Two topics dominated the discussion. The participants were impressed by the efforts needed to safeguard the existing material on traditional carriers (processing – including the generation of metadata, need up to 8 times real time) and the clarity of the message that the reliance on new storage media such as CDs and DVDs for storage is a trap and cannot be recommended. The other topic under discussion was the nature of language archives. They can neither be pure archives that rely on storage, nor can they only focus on access facilitation. Due to the dynamic nature of the secondary material which has to be stored together with the primary material – the recordings, they have to take care of both aspects. However, it is common practice already that reliable storage is taken care of by computer centers that act as partners. Language archives seem to be in between the

two poles: classical archives and digital libraries. It was agreed, however, that the pure existence of the material is primary. With respect to the presentation of material a wide spectrum of solutions is possible, archives cannot deal with all.

Researcher Requirements

Helen Dry and Tony Aristar gave the first talk about the wishes of potential users from different research disciplines that are dependent on language resources as their basis and from people working on educational publications. According to them it is necessary to have one central catalogue to easily locate interesting resources crossing archive boundaries, to easily preview material to see whether it is exactly what was expected and to get easy and un-bureaucratic access permissions. They see the need for very simple access interfaces and a combination of metadata and content search that can incorporate classes of features such as from morphosyntax and phonology. It is obvious that many people don't only want to search and visualize, but to add annotations and commentary.
Further, they addressed the fact that cross-corpus or cross-archival search requires the mapping of different terminologies used by the different (groups of) researchers to registered concepts from ontologies or data category registries as they are intended in ISO or being created within GOLD. When new resources are created, one can already refer to a growing number of tools that link to ontologies such as OntoELAN and FIELD. For legacy resources, however, mapping files are required. They finished by making clear that smart search engines are required that make use of mappings, language profiles and annotation indexes. The community needs architectures to tackle these difficult problems and not just individual tools.

Next Peter Austin, David Nathan and Robert Munro explained their views of what is requested. Giving a useful answer is not easy since the linguistic community is not a uniform one. They describe the typical fieldworker as kind of lone wolf and all archiving aspects are completely irrelevant to them. The focus of their talk was very much on mobile data which is data that can be used for example by language communities since its presentation form is adapted to what these communities are used to. Tools supporting smart data mobilization require special development methods that are different from the traditional ways. Therefore, they proposed an ethnography of tool development that includes the interactions with the users. Only critical self-reflection will bring us forward to develop tools that meet the needs of the different communities.

Discussion
During the discussion it became obvious that different types of communities may be interested in the data and that in particular the language communities should get easy access. Different models were discussed. On the one side there is the brilliant idea of mobilized data that is specifically tuned for specific communities, but these views can hardly be provided by the archives since special skills are required to prepare such guided tours or exhibitions and it costs much time to create them. Some participants stressed that tool development for field linguists has undergone years of cross-fertilization. So there is much knowledge already of what is needed. It was agreed that metadata based access paths to resources will not be the preferred way for language community members to locate resources. On the other hand the archives should give simple access as demanded by the first speakers so that for example education-oriented persons could easily create special educational material. Distributing complete sub-corpora to communities will be a relevant option since Internet connections will remain slow in many places. Where suitable connections exist the archives should give immediate access options. In this respect the requirements presented in the first talk were largely supported. Archival contents must be connectable so that joint queries can be handled without the need to register again and again. Also content enrichment by users is a necessary extension, the stand-off principle helps to separate the contributions.

Ethical and Legal Aspects

Gary Holton and Heidi Johnson spoke about legal and ethical aspects in so far as they are related to the DELAMAN goals. While laws define what people are allowed to do, ethical rules tell people what they should do. Copyright defines who the copyright holders are and how resource sharing can be done legally, i.e. rights to copy, distribute, publish etc, but these definitions vary between countries. Archives are bound to the local rights although the creators and depositors may come from different rights systems. However, copyrights are legally treated as property, which can be assigned to others. Yet there is a big problem, since formulations of copyright and IPR apply to individuals and not to

traditional or collective knowledge, which is precisely the type of material DELAMAN archives typically store.
Due to this situation moral and ethical issues are much more relevant for the DELAMAN archives. However, there are very different views about what people should do and the views are changing over time. The presenters gave a set of radical claims: (1) Archivists have to respect restrictions placed upon resources by creators and depositors for obvious reasons even if the views on restrictions are still in dispute. It is also obvious that we may not ignore those resources where restrictions are defined or where they are not at all clear, since we would exclude many language communities. (2) Archives must not allow local legal restrictions to inhibit preservation of and access to the world's linguistic heritage. Therefore, we have to make sure that original specifications about access management and control will travel with the resources without any exception. (3) Archives have to define the terms of the rights management issue (problem space, tools, education).
According to the presenters DELAMAN should agree on the types of access restrictions and the types of access controllers and work out the P2P implications for the intended file sharing.

Discussion
It was agreed that whenever we speak about access rights actually we are speaking about access policies that include procedures and statements that have to be assigned. The requirements of the archives have to be worked out. It was not clear what the level of granularity should be for granting access (fragments, objects). It was also suggested to come to a commonly agreed code of conduct since the current rules of the different DELAMAN archives don't seem to differ so much. It was agreed that the legal implications of rights traveling with the shared resources have to be studied. Finally, it was obvious that agreements can only be achieved by involving the creators and depositors in the discussions, i.e. the required time to come to agreements may not be underestimated.

Archivist Requirements

We asked three language resource archives to present their approach and their architectures: AILLA, PARADISEC and DOBES/MPI. All of them have already some years of experience and collected a comparatively large amount of endangered language and music resources. This report is not meant to give a detailed description, but to point to a couple of key characteristics.

AILLA
The AILLA archive was presented by Heidi Johnson. It was started in 2000 and is based on research in anthropology and linguistics. Technically it is supported by the Digital Library Services of the university of Austin. AILLA only houses digital language resources in more than 40 languages from 8 countries. The preservation of the material and lending access to it are the two pillars of AILLA's mission. Internet is the only way to access it and for the future this makes sense, since Internet cafes are springing up all over Latin America. The goal for the future is to establish a network of related archives incorporating regional centers in the language communities.
Heidi further described the importance of a suitable access to the archive material by members of the indigenous communities for language maintenance and revitalization and for researchers to carry out cross-language projects. Giving access is risky for religious, cultural and political reasons. Therefore, AILLA has worked out a graded access system covering four steps: free public access, automatic control, depositor control and indigenous control. These levels can be set for every individual file in the archive. Timers are set so that restrictions and passwords can be reviewed after a while. Experience shows that depositor and community control is very difficult to realize. In particular indigenous control leads to community disputes that are difficult to handle for an archive. Based on these experiences AILLA wants to work towards a distributed access management system where the archivist takes a neutral position and only gets notified about changes.

PARADISEC
The PARADISEC (P) concept was presented by Linda Barwick. It is a collaboration of four Australian universities with the intention of digitizing, safeguarding and making accessible endangered languages and music material from the Asia-Pacific region. Each of the four universities (Sydney, ANU, Melbourne, New England) has a specific role. The digitization and workflow processing is done in Sydney and the data is stored at the APAC national computer facility at ANU. The fast Australian Grangenet research and education network is used to exchange the bulk data. Melbourne is advisor in metadata and other data processing aspects. Much effort was invested to work out streamlined processes not only to salvage important recordings, but also to make them available to the interested community in a timely and cost-effective way. One important part of P's strategy is to return materials

to the language communities and ways for doing this are investigated with high priority. Therefore, the archive is focusing on excellent links with regional data centers.
Currently, the archive covers 1623 metadata records of which about 800 objects have been digitized with a repository size about 1 TB. The recordings come from 24 countries in the Asia-Pacific region. OLAC is used as metadata standard and metadata from the repository is exported regularly so that it can be harvested by the OLAC OAI-PMH harvester. Digitization using the Quadriga audio archiving system produces 48/96 kHz / 24 bit Broadcast Wave Format files, For access purposes MP3 derivatives are generated by batch processing. Online copies are made available via password-protected access to the APAC national storage facility and off-line tape copies are generated at Sydney and APAC. Costs are an issue for PARADISEC and other Australian users, for example both international data traffic and data exchange with Australian Universities not yet connected via the fast Grangenet backbone incur charges, and some Universities also apply internal data maintenance costs on a per-GB basis.
Access to the material via streaming options, suitable access management protocols and a geographic search interface are all in development for 2005. Finally, Linda Barwick mentioned some issues that have to be tackled within DELAMAN for it to succeed.

DOBES/MPI
The MPI is housing two archives: (1) the one that contains all contributions from its own researchers and (2) the one that is created within the DOBES programme which is funded by the VolkswagenFoundation and now covers 24 documentation teams. Both archives are part of an integrated metadata and organizational domain that currently contains about 40.000 sessions or bundles which means more than 100.000 language resource objects. Due to the digitized media streams the archives contain about 11 TB of data. Since for the DOBES archive more specific requirements with respect to format and organization were defined, it was taken as example for the talk.
The DOBES archive has as principle to take almost all digital resources that are in digital format or that can be digitized independent of their format. However, only a few formats, representing structured texts such as annotations and lexica, less-structures texts, images, audios and videos are accepted as archival formats. Further, all resources are described and organized with the help of the XML-based distributed IMDI infrastructure. Therefore, much conversion and formal checking work is being carried out with the goal to come to a coherent and consistent archive where at least the metadata is open to everyone. This format coherence allows the development of a couple of tools around the archive. Finished tools are (1) a complete metadata infrastructure making IMDI metadata searchable under Google, (2) an elaborate and efficient Access Management System and (3) basic methods to access the resources including media streaming. Currently, we are working on (4) a Language Archive Management and Upload System to open up archive manipulation and extension to users, nevertheless guaranteeing a high degree of archive coherence and consistency, (5) a modular web-based archive exploration framework including components to visualize and manipulate annotated multimedia resources and lexica and (6) a web-based framework for commentary on archival contents and collaboration.
The archive has a 2-layer architecture where the users and depositors just see the metadata layer, the system managers the physical structure of the archive and where the archive manager looks at both and takes care of consistency. This distinction allows the physical structure to be changed without notice to the users, for example when for example the HSM components are migrated to new storage generations. It also allows users to build their own virtual temporary structures and it opens the way towards Grid type of infrastructures. IMDI domains can be easily integrated to one browsing and searching domain and it is very simple to setup portals that integrate distributed resources as it was created in several large projects. Due to the architecture, the DOBES resources can easily be exchanged with other repositories. Two complete and dynamic copies are already maintained at two German computer centers not only for backup purposes.
The MPI/DOBES archives can easily be adapted to the DELAMAN/DAM-LR kind of infrastructures. Unique resource identifiers can be introduced, sharing of resources can be extended and the current access management system can be replaced by something else if there is mature and efficient technology that supports the requirements. What is definitively needed is robust and stable technology and a fall-back strategy, i.e. whenever we integrate our archives with others we have to make sure that we can give services when the integrated ones would be out of operation. The IMDI structure and semantics is obviously something that cannot be replaced, since they are result of several years of discussions with the community.

Panel Discussion
by Sven Stromquist, Marcus Uneson and Daan Broeder
In the panel a couple of issues about archive exploitation, archive building and other points were confirmed:

- users want to combine search on metadata and on content – the latter is important
- users want to search across different terminologies, i.e. cross-corpora
- it would be interesting to have a usage tracking mechanism (who is using the data)
- the stability of references to resources has to be guaranteed, URIDs may help here
- easiness of access to archival material is important, however, different groups have different views on what easiness means; one registration effort and one user ID is a basic requirement
- archives have to provide means to support local operation, i.e. whole sub-corpora should be object of distribution and then be accessible in the same way
- archives have to be dynamic, i.e. the ingestion of new resources and annotation layers and the improvement of existing resources should be simple
- access to fragments is seen as an issue (for example the first 2 minutes of a long recording), but the overhead may not be underestimated; if annotation tiers are in stand-off format then different access rights per tier can be easily maintained
- the task of an archive was brought up again; a pure preservation archive in the language resource area is doubtful; however, the cost aspect may not be underestimated if an archive does more than long-term preservation; this aspect is of particular relevance when funding periods stop
- in this context the question was raised whether for example the DOBES archive will carry on to accept new resources or updates[5]
- with respect to the exchange of the data between archives it was stressed that most important is a trustful relationship
- with respect to the ethnography of tools it was questioned whether there is a reference to an existing framework that is flexible enough to adapt to all types of users as a feasibility case
- the DELAMAN archives have to help each other in getting costless support for long-term preservation at computer centers and data exchange (this seems to be a high-priority topic)

Technologies and Products

Technological Overview
Bernhard Neumair, Thomas Soddemann and Egon Verharen shared the preparation of a talk that was meant to give a first overview about technical aspects that play a role when speaking about DELAMAN like goals.

Bernard Neumair first gave his view on network and storage capacity development. Basically he referred to the √2 rule which says that bandwidths and capacity are increasing by a factor two in two years. In this development the cost/bandwidth and cost/capacity ratio is continuously decreasing, i.e. the next generation is available for similar costs. High bandwidth networks are available so that even the transmission of audio and video is no problem. More efficient and cost-effective access points for audio and video transmission (DiffServ, MPLS) are technologically ready and will be made available soon. In the area of storage the √2 rule will be broken when completely new technologies will be available. He gave the example of holographic storage which from today's viewpoint will bring almost unlimited storage capacities. The introduction of this technology is not so far away anymore, a 1 TB DVD is almost ready to be announced. Very important for physical storage management is a reliable and intelligent HSM software system. It is obvious that smart media migration and refreshment strategies have to be implemented and that time has to be reserved to transfer the data.

Another key technology in the web area has to do with trust relationships, i.e. when centers are exchanging sensitive data we need to be sure that a service is indeed the one that it claims to be. This problem can only be solved when we come to an infrastructure where servers and services have to authenticate themselves. The setup of a Public Key Infrastructure (PKI) is a basis for securely encrypting all sorts of messages. It is based on two keys that are different but also related. To ensure confidentiality of your message you ask your partners to encrypt the message with your public key and you decrypt the message with your own non-public private key. To guarantee integrity you encrypt the fingerprint of your message with your private key and others then can decrypt this with the

[5] the DOBES archive will be dynamic also after the funding period and accept contributions from other contributors

public key. How can we guarantee authenticity, i.e. how can we make sure that the service behind a public key is the one it claims to be? This is done by asking for certificates where a trustworthy third party signs the relation between public key, name, organization, email address etc. Such certificates are created within hierarchical structures where parent authorities certify child authorities starting from one generally accepted root. Setting up and maintaining such a PKI system means overhead and is related with costs. There are efforts to come to an authority tree for the research world up to the top in Europe.

Thomas Soddemann then introduced the key concepts of Web Services (WS) as a generic technology to offer arbitrary services in the web. He first sketched the typical 2-tier and 3 tier client-server architectures where clients communicate with HTML web-servers or specialized servers that interface to enterprise applications and legacy databases. In contrast to this scenario WS are characterized by discoverable descriptions, an interface specification telling exactly what type of information the service needs and which information it can deliver (WSDL) and how messages are exchanged (SOAP). All is based on XML as underlying syntax. Due to its machine processable nature and its dynamic binding capability, WS can be combined to new services and this opens up a completely new degree of flexibility. WS allow to realize complex service-oriented architectures in the web. However, to realize for example a full shopping cart function one needs state information which is not provided in standard web-services. WS-RF (WS-Resource Framework), designed by the Grid community, has this feature. For more details on WS see below.

Finally, Egon Verharen gave an impression about a few new application concepts. In analogy to the layered OSI model he gave a layered hour-glass model for web-based application scenarios. The IP protocol is the basis of all interaction. On top of that he defined the lower middleware layer (searching, personalization, agents, URIDs, access management, etc). The upper middleware layer is defined by streaming, collaboration tools, metadata, webservices, digital rights management etc). These middleware components are being used by research and education specific tools. He briefly touched a number of applications such as capturing, editing, archiving etc that are involved in the lifetime cycle of video resources and gave an impression of the media streaming archive implemented at SURFnet. Special measures have do be taken to do authorization when video streams are requested. Finally, the example of a collaborative environment was presented that covers synchronous and asynchronous collaboration channels.
His final message was that everything DELAMAN wants to achieve can be built now. Integration and standardization are the primary focus which means making good choices.

Discussion
The costs aspect was raised that may occur when large amounts of archival data will be exchanged. The broad discussion in the whole Max-Planck-Society about long-term preservation resulted amongst others in the statement that the current amount of data can almost be neglected when a new storage generation will be installed. In general it will imply a factor of 10 times more storage capacity at the same price. Not all computer centers seem to follow that strategy, i.e. for some DELAMAN parties storage costs will play a significant role. Another aspect was what the perspectives are for a unified certificate authority tree covering all DELAMAN and/or DAM-LR archives. This point could not be fully clarified.
A discussion clarified that Web Services are a web-capable form of APIs and that most of us should not have to deal with this basic technology. However, it is good for all of us to understand the implications and relevance of web services.

Web Services
Thomas Soddemann elaborated on Web Services (WS) in greater detail, since they will form the basis of future web-oriented component software architectures. They have a standard interface definition language and a standard invocation mechanism. A few examples of well-known web services such as the Google Web API were given, i.e. such services could be integrated into more complex services.
Web Service providers have to publish their service in a machine readable way and register it at a service broker site. Here the UDDI metadata description is one of the suggestions to describe the services. However, UDDI is only a formal framework: the communities have to define the descriptors and organization. Service consumers can use the UDDI descriptions to find suitable services. Once found, the consumer has to bind to the WSDL-based interface specification of the service.
Web Services come along with a security architecture to encrypt the messages that are exchanged. It provides mechanisms to guarantee end-to-end integrity and confidentiality of messages. This can be

tuned such that different parts of messages can be made readable by different user groups. Since Web Services are stateless, WS-RF (Resource Framework) was designed by the Grid community. Keeping state information as foreseen in WS-RF would allow to build services that include for example a shopping cart application.

Grid Components

Dave Berry then spoke about the components and tools that were worked out in the Grid community. He referred to work within (1) the EGEE (Enabling Grids for eScience in Europe) project that aims to build on recent advances in Grid technology and that develops a service Grid infrastructure in Europe which is available to scientists 24 hours-a-day, (2) the GGF OGSA working group working on the Open Grid Service Architecture and (3) examples from other institutions building on GRID components. He pointed for example to the Digital Curation Center project in the UK that has goals comparable to DELAMAN/DAM-LR. The Digital Curation Center is a network of collaborating data organizations. While curation stands for distribution over time, Grid computing stands for distribution over space. Distribution over space makes it possible to create virtual organizations where people from different institutions share processing and data resources and collaborate. The Grid technology is meant to support these virtual organizations.

The main Grid topics are authentication & authorization, compute job submission, data replication management, logging & bookkeeping and monitoring. As an example of a Grid middleware system, the Globus Toolkit version 2 has corresponding key components: Grid Security Infrastructure (GSI), Grid Resource Allocation Management (GRAM), GRID-FTP, Grid Resource Information Service (GRIS) and Monitoring and Discovery Service (MDS). The Open Grid Services Architecture (OGSA) will describe a complete architecture using Web Services and WS-RF to provide the above mentioned goals. Many of the Grid ideas emerged from the requirements of computer Grids where powerful compute servers work together to handle large compute intensive tasks, but the project called OGSA-DAI (Data Access and Integration Services) deals with the specific problems of sharing data. Registries with metadata information are involved to find information about the location of data resources and access aspects have to be dealt with. There are a few implementations of Grid components based on Web Services: Globus Toolkit version 4, GLite (being worked out in the EGEE project) and OMII by the Open Middleware Infrastructure Institute. In particular, GLite has provisions for file replication services. Finally, Dave added that there are still a number of open questions including the naming system, metadata standards and various architectural details.

Authentication and Authorization

Bart Kerver spoke about requirements and solutions for comprehensive Authentication and Authorization Infrastructures (AAI). The main motivation for AAIs can be found in personalized service provisioning, educational and network mobility and the need to reduce the amount of digital keys people have to handle. Currently, every web application implements its own user authentication, however Bart voted for splitting this and let specialized services do the authentication. Different levels of security could be handled by such specialized services ranging from a username/password mechanism to the check of biometric information. He mentioned a few authentication systems that provide various levels, one is the A-Select solution developed by SURFnet. Authorization is another service that will clarify whether a certain identified user is allowed to access resources. Again he advocated using special services that have policy engines supporting attributes such as roles. These also can support complex privilege hierarchies.

Federations, organizations that share data and access management, have to apply cross-domain AA, i.e. agree on policies (procedures, agreements, etc) and technologies (protocols, schemas, PKI, etc). Shibboleth is a software product for federations offering authorization, attribute gathering and safe transport of attributes. However, it does not do authentication. He showed how A-Select and Shibboleth can cooperate in a federation environment to form a complete AA infrastructure. It seems that Shibboleth will be widely used in academic federation scenarios, obviously in combination with various authentication solutions. Some convergence to standards and an implementation of a single sign-on option across networks will be necessary to realize the DELAMAN AA infrastructure.

Handle System

Larry Lannom reported about the need for persistent resource identifiers and described the Handle System that can resolve persistent identifiers to physical paths in an efficient way. Before introducing this topic Larry made some words about the experiences from the discussions in the Digital Library community. According to this language resource archives fall in between the classical archives and the new digital libraries since they have to take care of both aspects: store data reliably and provide

interactive access to the material. However, the language resource community is much less heterogeneous than the DL community, i.e. good data and usage models can and should be developed that serve as basis for all DELAMAN activities.
The Handle system is an efficient resolution system that resolves names into attributes. This allows us to talk with abstract names and not care about location details etc. It could be used for example to resolve unique resource identifiers into physical paths. The introduction is often a must, since physical paths (URLs) are very fragile. Locations are changing due to migration and distribution policies. However, such a resolver is another layer of complexity and it has to be administered. The HS allows associating handles with types, offers a distributed, scalable and secure framework and is optimized for speed and reliability. A number of key users are already making use of HS including the Library of Congress, the scholarly publishing industry, and it is being integrated into the Globus Toolkit.
A handle is a dual number: one part - the prefix – specifies the registered domain and the second part is a number generated by the domain. Each registered domain is free to setup its own numbering system. The prefix is stored at a Global Handle Registry (GHR) so that resolving of any HS conform identifier is guaranteed. The typical request is that someone wants to know all paths that belong to a certain handle. The Handle System is a collection of handle services that can have replicated sites each of which may have several servers. This architecture offers redundancy to guarantee availability and efficiency to guarantee speedy services. Local handle resolving can be done without contacting the GHR which prevents access bottlenecks. HS comes with an administration tool to administer handles.
HS users have to make choices about what they want to identify (abstractions, manifestations) and have to decide whether they want to associate more information such as metadata with a handle. Actual applications show that HS can handle many requests with little overhead.

Storage Resource Broker
Finally, Reagan Moore presented the Storage Resource Broker (SRB) which is a comprehensive approach to support federated archives as it is intended for example in DELAMAN. It seems to be the most far developed system to support data migration, data exchange, shared collections, simplified access options and federated server architectures supporting distributed data. It can be seen as a shell on top of local archives and therefore comes close to the visions discussed within DELAMAN and it is already widely used in similar scenarios, i.e. it has proven its reliability.
In this summary we can only indicate some of the highlights of the SRB package. It allows developers to build shared collections where users are authenticated, access is controlled and the file name space is organized, all independently of the storage systems. Further, the metadata layer is managed independently of the content and measures are taken to maintain consistency. Abstractions are introduced and maintained where this is necessary to support archive federations. This allows archives to register their resources under SRB without consequences for them. While registering, unique identifiers and metadata descriptions representing the resources are created. Metadata can be organized hierarchically in SRB and any sub-collection up to a single resource can have its own metadata attributes[6]. SRB allows users to browse and search in these hierarchical metadata domains. Important for the usage of the registered resources is a single sign on which is implemented by introducing a logical user name space.
Elaborate techniques are available to support fast and reliable data exchange, including parallel streams and checksums and the copying of access control information with the replications of the resources. Synchronization between copies can be taken care of. To handle large data volumes so-called bulk operations are provided. SRB also includes some basic services such as searching across all storage systems, user defined hierarchies and multiple ways of discovery which can be metadata or based on unique identifiers. SRB also offers a number of interfaces according to digital library standards such as OAI-PMH and it supports modern API standards such as WSDL. The code distribution policy is that the code is available to academic institutions, however, there is also a commercial version.

Panel Discussion
The groups of linguists and archivists were asked to reflect on the presentations. Several of the points that came up were already presented in the workshop results and will not be repeated again. There were a couple of additional technical aspects which will be listed in short:

[6] The new version 3.3 being launched these days will give full support for user-defined metadata sets.

- It was stated that the Grid community is a rather heterogeneous community. The answers to questions are dependent on whom one is talking to. For people not involved in the Grid developments the state of the components is very unclear and the documentation is difficult to interpret.
- It was also stated that the Grid community makes some decisions that bring them into competition with other international standards (MDS versus UDDI[7]). For DELAMAN this is important since the primary focus is not on joining compute servers which was (and perhaps is) the major focus of the Grid people.
- SRB is a highly interesting product, however, with respect to the registration of data some work has to be carried out by the data providers such as for example a metadata mapping. The current metadata model is flat and does not support structure, i.e. the metadata descriptions are a list of attributes. SRB offers APIs, but the programming has to be done by the contributors.
- The SRB licensing policy was a point of concern, since language archives work on small budgets. Expenses for software are very problematic and there is a fear that license costs could come when software has got a certain market share. The availability of Source Code is nice, but not sufficient due to code maintenance aspects.
- It was explained that there is much more data modeling and international standardization on the way in the language resource domain than may have appeared from the workshop. The reason for this gap in information is that this workshop was not meant to repeat all the discussions about abstract metadata, annotation, lexicon and other models, since this would have taken too much time.

The reflection discussion was rather controversial nevertheless highly interesting. We will restrict ourselves in this report to mention a couple of major statements that were made and where further debates will be necessary:

- A lack of a shared universe of discourse was noted between the workshop participants who come from very different disciplines. Technologists are in general handling Vanilla data while linguists were not ready to specify their real needs.
- The discussion about rights management, IPRs and copyright were not elaborated enough. It was questioned whether linguists were given sufficient context to be able to make proposals that technologists can work with.
- Archive resource replication is a primary focus, but there were no statements yet whether linguists for example like it. Aspects of control about such a process are not yet clear.
- The term user was used in a misleading and unclear way. It is not yet clear which the user groups will be and therefore statements about user interfaces were vague. In particular the needs of the language communities are largely unknown. Therefore, a step backwards was suggested to collect user feedback about developments and to better understand the needs.
- It was suggested that instead of presenting prototypes of tools to the users they should be part of a much more abstract process of defining project needs, core principles, contexts etc from the very beginning.
- The notion of the basic data object is not quite clear. Is it an annotation resource, an annotation layer or even an annotation fragment? Depending on the answer the implementation of access management for example may vary.
- It was questioned whether there is enough knowledge about each others discipline (linguistics, technology) to expect a productive interaction[8].

References/Glossary

AAI	Authentication and Authorization Infrastructure
AILLA	www.ailla.utexas.org/site/welcome.html
ANNEX	Web-based Annotation Exploitation Tool (available soon from MPI)
API	Application Programming Interface
A-Select	a-select.surfnet.nl/

[7] The main difference is that MDS supports dynamic services, i.e. creation and deletion over a relatively short period.

[8] It should be added that the majority of participants did not share this skeptical view.

DAM-LR	Distributed Access Management for Language Resources (EU Project – web-site to come)
DELAMAN	www.delaman.org/
DOBES	www.mpi.nl/DOBES
Digital Curation Centre	www.dcc.ac.uk
EGEE	public.eu-egee.org/
ELAR	www.hrelp.org/
ELAN	www.mpi.nl/tools/elan.html
OntoELAN	csdl.computer.org/comp/proceedings/ismse/2004/2217/00/22170329abs.htm
E-MELD	emeld.org/
FIELD	www.ling.ed.ac.uk/linguist/issues/13/13-2510.html
GGF	www.Gridforum.org/
GOLD	emeld.org/gold-ns/
GTK (Globus Toolkit)	www.gtk.org/
HS (Handle System)	www.handle.net/
HSM	Hierarchical Storage Management System
IASA	www.iasa.org/
IMDI	www.mpi.nl/IMDI
IPR	Intellectual Property Rights
ISO TC37/SC4	www.tc37sc4.org/
JPEG	www.jpeg.org/
LACITO	lacito.vjf.cnrs.fr/
LEXUS	Web-based Lexicon Tool (available soon from MPI)
LREC	www.lrec-conf.org/
MDS	www-106.ibm.com/developerworks/Grid/library/gr-mds.html
MPEG	www.mpeg.org/
MPI	www.mpi.nl
OMII	www.omii.ac.uk/
OGSA	https://forge.gridforum.org/projects/ogsa-wg/
OGSA-DAI	www.ogsadai.org.uk
OLAC	www.language-archives.org/
PARADISEC	paradisec.org.au/
PKI	Public Key Infrastructure
Shibboleth	shibboleth.internet2.edu/
SOAP	www.w3.org/2002/ws/
SRB	www.npaci.edu/DICE/SRB/
TIFF	home.earthlink.net/~ritter/tiff/
UDDI	www.w3.org/2002/ws/
UNICODE	www.unicode.org/
URID	Unique Resource Indentifier
VolkswagenFoundation	www.volkswagen-stiftung.de/
WAV	www-ccrma.stanford.edu/CCRMA/Courses/422/projects/WaveFormat/
Web Services	www.w3.org/2002/ws/
WSDL	www.w3.org/2002/ws/
WS-RF	www.globus.org/wsrf/
XML	www.w3.org/XML/

Acknowledgements

We would like to thank the VolkswagenFoundation for having funded this expert meeting and the MPI for providing its facilities.