DAM-LR

Distributed Access Management for Language Resources

Summer Workshop

Concepts and Technologies for the Integration of Language Resource Archives

University of Lund, 21-23. August 2006

The EC-funded DAM-LR project is currently busy in developing components for an integrated domain of language resource archives, to implement a complete architecture and to define the formal basics of a federation of archives. Language resource archives have requirements that are more related to the Humanities' needs than to those of the hard sciences where the focus is on compute Grids to tackle the Grand Challenges. This is reflected in the type of technological solutions and the organizational framework for a federation that have been chosen. With respect to all aspects intensive discussions have taken place. In particular, with respect to the technological aspects solutions and components from the domains of Grid initiatives and Digital Library initiatives have been investigated.

For the workshop the DAM-LR partners defined the following goals:

- discuss the tasks and principles of language resource archives (***Live Archives)***
- inform other interested parties about federation technology and useful components
- explain the chosen architecture and discuss additional wishes and possible problems
- give a more detailed insight into the functioning and requirements of the major components
- discuss the organizational requirements to establish a federation of archives
- discuss the possibilities and requirements to broaden the current small DAM-LR project towards an integrated European federation of language resource archives

The intention is that the participants will get a good overview about possible solutions to establish a federation of archives, both technologically and organizationally, and to get a clear picture about the required investments when participating in such a federation.

With respect to technology we will elaborate on a number of essential components such as trusted certificates, Public Key Infrastructure, joint metadata infrastructure, open archive systems, the Handle System, OpenLDAP, Shibboleth and an extended Apache web-server as resource manager. Some time will be reserved to discuss component-based versus complete solutions. The presentations will be given by experts from within the DAM-LR project and where necessary from external experts.

There will be no fee for the course itself.
Registrations can be done via the web-site
www.mpi.nl/dam-lr
Due to practical sessions there is only limited space available.

DAM-LR: Distributed Access Management for Language Resources

Europe is home to many collections of linguistic resources, but each archive tends to use its own indexing and cataloguing systems, many of which are not even available on-line. Systematic searches across archives seem all but impossible. With the help of more than €380 000 in EU-funding, the DAM-LR project aims to create a single, virtual linguistics resource. Four partner institutions will use a common indexing framework and system interfaces so that their digitised archives can easily be integrated into a distributed system. The ability to access multiple archives seamlessly should improve the efficiency and effectiveness of linguistics research, and help Europe keep its historical dominance in this field.

Integrated access to linguistics archives

Europe has a rich linguistic and cultural heritage that has fascinated scholars for hundreds of years. Consequently, the continent has become a world leader in linguistics research and is home to a number of internationally important linguistics resources, such as multimedia archives and lexicons.

But while researchers are fortunate to have such a wealth of source material available for analysis, it is not always easy to access. Archives are located in institutions across Europe and they are indexed in a variety of ways. A few you can browse over the internet, but many can only be accessed by visiting the repository.

Several linguistics centres in Europe are working on ways to structure, catalogue and digitise their collections, and make them available on-line. Together they are creating indexing taxonomies and using a common framework of keywords ('metadata' tags), to label resources within their archives. This form of indexing allows researchers to find and retrieve relevant material quickly and easily.

The DAM-LR project, which is funded as part of the Sixth Framework Programme's Research Infrastructures action, brings together four linguistics research institutions from the Netherlands, Sweden and the United Kingdom. The partners are able to share their expertise and experiences and, most importantly, collaborate on interoperable systems. These could then be integrated together in a way that would allow researchers to access the archives of all four participants seamlessly, as if it were one single collection.

Virtual integration

Integration of extremely diverse archives is possible because each partner's locally adapted system has to meet set specifications. The systems in development at each of the four partner institutions share:

- a common metadata framework, called the ISLE MetaData Initiative system (IMDI);
- the use of unique codes for each archived item (known as 'unique resource identifiers');
- a common system that could allow researchers, ultimately, to access all of the resources with a single user profile; and
- an access management system that would allow archive managers to set user access rights.

The IMDI metadata is already used by about 50 institutions worldwide, including European language resource centres and sign language communities in many European countries.

A prototype management system, called the Language Archive Management and Upload System (LAMUS), has also been developed and the four partners are using it to build their local archive systems. Probably the most advanced is the digital archive at the Max-Planck-Institute for Psycholinguistics. It now holds about 40 000 so-called 'bundles' of material that cover over 100 000 individual items, such as audio and video media, linguistic annotations, lexica, field notes, sketch grammars and documentation. Researchers from different projects are able to add and integrate new data to the existing collection, which now amounts to about 11 Terabytes of data.

Once these local systems are in place, the remaining task will be to integrate them so that researchers can visit and browse the archives – perhaps without even leaving their desks. The benefits of a single, virtual 'super-archive' for Europe cannot be ignored: linguistics researchers will be able to find the best resources for their work more quickly and without the need for costly trips to visit archives in person. The quality of linguistics research stands to improve, helping Europe to maintain its historical lead in this field.

Distributed Access Management for Language Resources, in summary

Project acronym: DAM-LR

Funding scheme: Construction of New Infrastructures (CNI)

EU financial contribution: €382 000

EU project officer: Maria Theofilatou

Duration: 36 months

Start date: 1 January 2005

Completion date: 31 December 2007

Project webpage: http://www.mpi.nl/dam-lr/

Coordinator: Peter Wittenburg, Max-Planck-Institute for Psycholinguistics, The Netherlands

Partners: Max-Planck-Institute for Psycholinguistics (DE), Linguistics Department University of Lund (SE), School of Oriental and African Studies, University of London (UK), Institute for Dutch Lexicology (NL)