1. Why is the Humanities lagging behind?

1.1. The cultural heritage

A central concept in the Humanities is the “Cultural Heritage” and a lot of money has been spent over the past decades to make that heritage available on the Internet. This is not a small task, however, since the cultural heritage encompasses more or less anything which is man-made or touched by man. According to a definition formulated by the World Bank at its meeting in Washington 1988, “Cultural Heritage encompasses material culture, in the form of objects, structures, sites, landscapes, as well as living (or expressive) culture as evidenced in forms such as music, crafts, performing arts, literature, oral tradition and language. The emphasis is on cultural continuity from the past, through the present and into future, with the recognition that culture is organic and evolving.” It is not just the size of this pet object in the Humanities which makes it hard to put it on the Internet. Perhaps more importantly, its diversity poses a problem for its searchability on the net, since searchability presupposes metadata and metadata presupposes a consensus over ontologies. The project European Cultural Heritage Online (ECHO) states in its report [1] of the state-of-art concerning putting the cultural heritage on the Internet, that “…part of the explanation for the lack of accessible digital content in the humanities so far has to do with lack of metadata, infrastructure and interoperability, as well as with insufficient attention to user perspectives.”.

The ECHO state-of-art report [1], which targetted four areas in the Humanities – Languages, Ethnography and museum objects, History of science, and History of art – further stressed “…the importance of sharing resources as a means to arrive at a more creative and efficient usage of data and, thereby, to enhance the vitality of the cultural heritage. This process will have a positive effect already within disciplines, but the great potential lies in its application across disciplines.”. The report [1] further discusses what looks like an asssymetry between activities typically conducted by Research groups (RG) and Cultural institutions (CI) such as museums or libraries with respect to research data at their disposal. The situation is summarized in Table 1.

<table>
<thead>
<tr>
<th>Collecting data</th>
<th>RG</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematizing data</td>
<td>RG</td>
<td>CI</td>
</tr>
<tr>
<td>Making them accessible</td>
<td>CI</td>
<td></td>
</tr>
<tr>
<td>Conducting research</td>
<td>RG</td>
<td>?</td>
</tr>
<tr>
<td>Conducting education</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 1. Target activities in Research groups (RG) and Cultural Institutions (CI)

In short, cultural institutions are good at archiving data and making them accessible, but their data are underused for research purposes. In contrast, research groups conduct research on the data they have collected, but they tend not to make their research data available to users outside the research group. Neither of the two (RGs or CIs) tend to conduct educational activities in relation to their data. As we shall argue, the last type of activity is crucial for the advancement of e-science in the Humanities.

1.2. Language resources

In a project much inspired by ECHO, Distributed Access Management of Language Resources (DAM-LR) [2], efforts were concentrated to the linguistic sciences, in order to forward technologies and frameworks of cooperation to implement searchability and sharing of research data on the Internet. The linguistic sciences is a field within the Humanities which has already established a strong tradition of computational tools as well as the world wide web. With the breakthrough of personal computers in the early 1980s, computerized text corpora and associated tools for analysing these corpora became standard components of the linguist’s tool box. Today, this tool box comprises multimedia, and the text data are linked with audio and video data. Further, there is a growing body of research data from laboratory environments where language activities such as reading and writing are studied as they evolve in real time, comprising eye-tracking data and key-stroke loggings. It was therefore a strategical choice by DAM-LR to focus on linguistic sciences where there is a well established understanding of computer tools and information technology.
Despite the technological development in the linguistic sciences, however, there is still strikingly little sharing of research data. There are many research groups who have collected very interesting language samples, digitized the material and developed interesting analysis tools. But these efforts are seldom coordinated and the resultant digitized materials are seldom shared with other research teams. Often, the software developed is not generalizable/applicable to other digitized data than those of the local group and the particular programmer may well be the only person with proper knowledge about the program developed. The situation means that a lot of efforts and ingenuity is never propagated outside the local group and thus represents a considerable waste of resources. Interoperability issues as well as incentives for daring to share both data and methods across research groups should therefore be a top priority.

The framework developed by DAM-LR uses PMID as a metadata format: PMID is primarily tailored to bundles of research data with recordings of linguistic activities at the core. It is sufficiently flexible, however, to accommodate research data from research projects in neighbouring sciences. DAM-LR was thus targeting linguistic sciences in an effort to create an exemplar which could then be followed and modified according to the needs of other branches of the Humanities.

In the new ESFRI program “Common Language Resources and Technology Infrastructure” (CLARIN) the combination of e-Humanities with language technology fostered in DAM-LR is made technologically more advanced and scaled up considerably. In March 2008, there are more than one hundred member institutions in 32 European countries attached to CLARIN. See [3].

The final workshop of DAM-LR was organized in Lund in December 2007 with the intention to inform in particular those institutions interested in CLARIN about the basic layer of research infrastructure developed in DAM-LR. 36 colleagues from all over Europe participated in the meeting and created a productive atmosphere. The first day of the workshop was devoted to information on a management level and the second to the details of the various technological components and solutions.

2. How can DAM-LR and CLARIN be integrated with research and education?

2.1. Infrastructure initiatives

Now, given the technological solutions and networking opportunities, how should we proceed to build a framework which is conducive to a culture of users and contributors to frameworks for e-Humanities, such as DAM-LR and CLARIN? Such a framework must be built by a combination of top-down and bottom-up approaches. In terms of the former, we need to identify resourceful institutions and organizations which are in a position to recognize the added value of investing in relevant infrastructure and sharing research data and analysis tools in the Humanities. A national research council is a good candidate for such an institution, and, indeed, there is a growing interest in supporting research infrastructure, including means for sharing research data. In that context, there is a special pressure on the Humanities to demonstrate their needs and potential added values, since the infrastructural initiatives tend to target natural science, medicine and engineering science, with some opportunities for social sciences but very little, so far, for the Humanities. Again, the leading role of linguistic sciences and language technology within the Humanities means a special responsibility for demonstrating the usefulness of infrastructural initiatives and e-science in the Humanities. CLARIN is a very encouraging step in this direction which has gained recognition on the European level (ESFRI) and which is gaining increasing recognition on a national level.

Another candidate for an institution which is in a position to recognize the added values of an initiative like ECHO or DAM-LR is a university. Making research data searchable and accessible has the potential of enhancing research, education and electronic information/library services – all of them central functional domains of a university. The university also offers a context for bottom-up approaches, insofar that there exists a number of practices which are closely affiliated with the concepts of ECHO or DAM-LR and which can be expanded and enriched by the functions of the DAM-LR framework.

For the purpose of illustration, I will briefly review – in section 2.2 below – the various ways in which we are trying to integrate e-science practices with existing research projects, educational programmes and library services at the recently created Centre for Languages and literature at Lund University. Among other things, we are aiming at creating a culture of e-science, where students and young researchers learn to create metadata for their own research data – just like they learn to create bibliographical metadata, for example, for their M.A. papers, in order to make them accessible in the data base system of the University library. Thereby, they learn to create and use a common resource of research data for the benefit of themselves and other researchers. The framework of the project DAM-LR is used as the technical basis for our educational project.

The core facility for the project is the Humanities laboratory at the Centre for languages and literature [4], a cutting edge laboratory for language, culture and cognition, with units for the empirical study of
behaviour online: an eye-tracking unit, a tactile unit, a body motion tracking unit, a phonetic-acoustic unit, an electrophysiological unit, a virtual reality unit, a unit for the study of writing online, and a unit for digitization and the ingestion of data. The lab has a core staff consisting of an eye-tracking expert, a methodologist, an audio/video/VR engineer, an expert on electrophysiological – especially ERP – experiments, a programmer, an IT-pedagogue, and a data manager. The policy of the lab is to offer methodological advice to senior researchers as well as masters and PhD students and to teach or guide them to use the lab facilities for implementing empirical research projects - smaller or larger ones. The personnel further gives regular courses in the Masters and PhD programmes, including statistics, experimental methodology, eyetracking, electrophysiological methods, corpus-linguistic methods, research ethics, and computational methods for research in the Humanities. The lab has a close cooperation with the technical group of the centre and with the research library of the centre. In January 2006, a metadata server was set up in the Lund lab in cooperation with Peter Wittenburg’s technical group at the Max Planck Institute for Psycholinguistics (MPI). In August 2006, a new 50 TB Language Archive/Data provision Server was purchased and is now operational. The Humanities lab and its network thus offers an infrastructural backbone for initiatives such as DAM-LR and CLARIN, not the least for helping implement a user culture in research contexts and educational contexts at the Faculty of Humanities at Lund University. More generally, the Lund initiative can be thought of as a test bed for implementing a user culture for e-Humanities.

2.2. Eight contexts for the integration of DAM-LR at Lund University

The first context is research. A variety of research groups at the Faculty of Humanities at Lund University were invited to seminars, demos and training events, where we explained the concept of transferring your research data to hard disk and organizing them by means of IMDI for the benefit of your own research group, other researcher groups and for the benefit of students trained in your field of research. The research groups were all asked to detail their needs, typical dilemmas, doubts and demands. A budding dialogue about new ways of organizing and sharing research data was thus initiated. So far, this process has led to the commitment of researchers from some ten different disciplines preparing local linguistic research data for access. These efforts include first and foremost four areas:

1. SWEDIA – a phonetic corpus of Swedish dialects, unprecedented in scope and detail
2. Swedish and Thai longitudinal child language corpora – approximately half a million running words each plus extensive video linkage
3. Archive of Kammu language and culture (sound and video recordings, photographs, drawings, music)
4. Recordings of visual and tactile reading online (eyetracking, finger tracking) and of writing activity online (key- and pen-strokes logging) – to be extended with gestures (body tracking)

The four areas are all cross-disciplinary. For example, area 3 involves researchers from Linguistics, Ethnography and Musicology, and area 4 researchers from Linguistics, Psychology, Cognitive science, Neuropsychology, Speech therapy, and Technical rehabilitation research.

The second context is the Humanities laboratory at Lund University, a new type of lab which produces masses of experimental data of language behaviour online. The production of this kind of research data is very resource demanding and it is therefore important to find an efficient way of sharing them, so that broader layers of the research community can make use of them.

The third context is education. We already provide support and organize local and international training events where the participants learn to digitize analogue research data, use IMDI to describe these data with metadata and organize corpora of empirical research data. The next step is to integrate these elements with standard courses given to students at levels where empirical research tasks are mandatory, most importantly the Masters level and the PhD level. Currently, the so-called Bologna process is being implemented across the Universities of Europe, with the goal of increasing the mobility of students and young researchers. A special effort is made to create attractive conditions for students at the Masters level, so that they can distribute their training across different countries. In this situation, DAM-LR offers a framework which means an important support for the increasingly mobile young scholar. Further, the growing content of the DAM-LR framework (see areas 1-4 above for an illustration) offers a rich material for lectures, courses and training sessions in linguistics, ethnography, musicology, cognitive science etc.

The fourth context is the library and its electronic information services. In general, the DAM-LR framework means a powerful extension to the modern research library, providing real, searchable research data. Among the electronic information services at
Lund university, there is a growing interest in customized interfaces for the individual students and researchers. Recently, a system called *My publications* was launched, where researchers and advanced students place electronic copies of their articles, book chapters, theses, course papers etc in the Lund Library server and classify them with bibliographical metadata in order to make them searchable. Similarly, a system called *My course library* was launched shortly afterwards. The next step is to create a customized electronic service, *My research data*, using the DAM-LR framework and its content as its basis.

The fifth context is publishing. Access to the research data underlying a publication – in combination with a careful description of how it was analysed - has several attractive potentials: a deeper understanding of how the results were reached, increased reliability etc. Offering this access as an additional or optional layer to the core publication is interesting for researchers, but also for students who can learn from the exemplar. Lund University has several publishing agents who can initiate this kind of publishing concept. A fortuitous circumstance here is the fact that the Lund University Board recently decided on a policy encouraging open access.

The sixth context is Public presentations. The DAM-LR framework and its content can serve as a useful source for presentations of research to the public in the context of museum exhibitions, web-sites, popular-science publications, teve programs etc. Recently, the Humanities laboratory at Lund University profited from the DAM-LR framework and its content for a regional project called *Exhibit Øresund*, where new ways of presenting research to the public are tried out [5].

The seventh context is International cooperation. The Humanities laboratory in Lund will profit from the DAM-LR framework to facilitate the sharing of online reading and writing data with several labs in Europe, such as LACO-CNRS, Poitiers. The lab has also a growing cooperation with the Wallenberg Global Learning Network (WGLN), an organization promoting among other things the usage of e-science in educational contexts targeting both universities and K-12 schools.

Finally (eight), the framework of DAM-LR contributes to an infrastructural advantage in the context of applications for research projects or the organization of educational programmes (e.g., Bologna). The eight contexts for the integration of DAM-LR at Lund University are summarized in figure 1.

---

![Diagram](http://www.xibit.info)  
Figure 1. Integrating DAM-LR with the Lund University context

### 3. Added values and the future

To sum up, the integration of the DAM-LR framework with various contexts at Lund University will:

- facilitate research cooperation
- enrich education
- facilitate mobility of researchers and students (e.g., the Bologna process)
- solve potential conflicts between cultural institutions and research groups
- enrich presentation projects (museums, encyclopedias, popular science, media)
- create an infrastructural advantage

We are starting with the linguistic sciences, but hope that the force of the exemplar will help extending the concept to neighbouring disciplines in the Humanities. To facilitate the creation of a broader user culture, we need, among other things, to build more user-friendly interfaces and to organize effective education and support. This process has to proceed in a creative dialogue with the users.

### 4. References


[4] [http://www.xibit.info](http://www.xibit.info)

[5] [http://www.xibit.info](http://www.xibit.info)