Appendix B: Experiences from DAM-LR

- a technological note -
- updated version -

Peter Wittenburg, Daan Broeder
DAM-LR Project
29.12.2007

1. Introduction

The DAM-LR project is one of the first Grid projects in the humanities - certainly at supra-national level[6]. Its task is to bring together archives with relevant language resources and allow users an integrated view on the resources they are interested in. On purpose, the four archives were chosen from different types of institutions and with different types of service focus. After having worked hard on the Grid implementation in 2006 we can draw first conclusions which will lead to a number of conclusions for the future.

2. Experiences

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

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

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

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

3. Conclusions

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

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