Appendix C: Language Archive Federations

Basis for Federation Agreements[7]
updated version by Peter Wittenburg[8]

Peter Wittenburg, Daan Broeder, David Nathan, Sven Strömqvist, Remco van Veenendaal
DAM-LR Project
25.12.2007

1. Introduction

The DAM-LR project can be seen as a test-bed for future infrastructures in the humanities and beyond to enable the eResearch paradigm. On a small scale it is a time limited project to establish a federation of language resource archives that share a Grid integration layer. It does so by testing out proper technologies that will allow us to do the required virtual integration. However, as a first paper about federation aspects (FFN) [1] pointed out, a federation is more than sharing a few technologies. It is also one of the tasks of DAM-LR to investigate these non-technical and primary aspects of "federations" especially in our domain. To prevent ending up in endless abstract discussions, however, we will do this based on the concrete needs of infrastructure projects such as DAM-LR and in particular CLARIN.

2. Types of Federations

When consulting Wikipedia for the term "federation" [2] we find the basic principles of state organizations which are the most "deep" domains where the term "federation" is used. We can read the following:

> A federation (Latin: *foedus*, covenant) is a union comprising a number of partially self-governing states or regions united by a central ("federal") government. In a federation, the self-governing status of the component states is typically constitutionally entrenched and may not be altered by a unilateral decision of the central government. The form of government or constitutional structure found in a federation is known as federalism (see also federalism as a political philosophy). It can be considered the opposite of another system, the unitary state

Such "Deep Federations" include detailed constitutional regulations that are ultimately broken down into legislative requirements that define constraints for example on the citizens. A federation is seen here as an alternative to a centrally organized state, since the members of such a federation retain some self-organizing power. There are many different examples for such federations that differ in the balance of power. The European Union can be seen as an example of a loose federation where the individual states retain much power and where the central government is comparatively weak.

In computational areas where very sensitive material is used such as in the medical domain, the virtual integration of data resources is also very much subject to very detailed regulations. So, for example, federations of hospitals have to establish an extensive set of rules of how to exchange and use patient data. In this case we can also speak about "Deep Federations".

Compared to such "Deep Federations" we can refer to a large number of "Shallow Federations" with much less detailed regulations. FFN speaks about the "Google-Federation" where participants restrict themselves to certain formats and principles so that their web content can be harvested to become indexed. There are even no explicit signed agreements, just a common understanding is sufficient to achieve world wide integration of open web content. All participants share the same common belief in the usefulness of world wide data mining, i.e., they share the same mission.

In the DSpace domain [3] users of the software discuss turning the user group into a federation that shares a number of interests that require minimal governance rules such as enabling communication

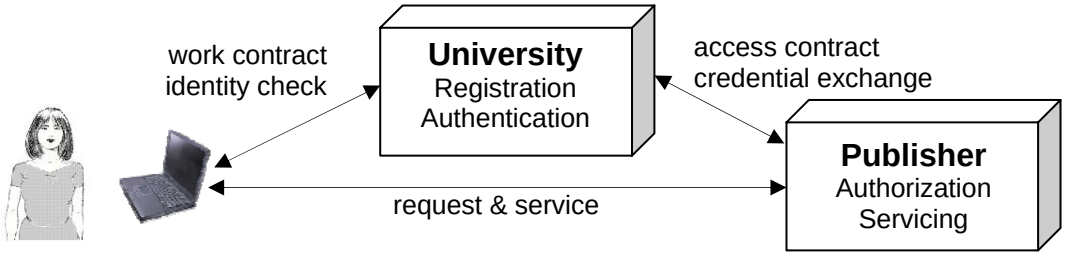
[7] This report is based on an earlier document created by David Nathan and Remco van Veenendaal which will be called First Federation Note (FFN).
[8] This report was updated by Peter Wittenburg, but the essentials were not changed.

within the DSpace community, ensuring that the community is healthy and resolving of conflicts is possible. In this case we can speak about a shared mission and a very loose definition of membership.

N. Volanis and J. Dumortier [4] distinguish between two models of Grid computing[9] and describe their legal basis. The **social model** "views the benefits of grid computing as a resource to be harnessed for the good of the society". Meeting the social model's objective - the achievement of the scientific goal - relies heavily on the moral value of helping the society by facilitating scientific research. The operational model depends on the voluntary submission of resources and in many cases the relationship between the partners is limited to the acceptance of terms of using given software. None of the actors engaged in the social model is willing to commit himself in a legally binding relationship that creates financial claims, obligations and responsibilities.

On the other hand, the **commercial model** sees in grid computing various business exploitation opportunities, i.e., companies need to control the resources to guarantee a Quality of Service. A number of enterprises can also form a Virtual Organization to share their data and resources based on a contractual relationship. These relationships will require severe financial constraints, controls and remedies, thus they require a "deep federation".

A kind of hybrid model is applied when for example large research institutions such as universities or groups of universities want to give their researchers access to a set of electronic versions of journals from publishers. The publishers will extend their normal set of regulations that define the usage of articles to the electronic domain and each user has to accept these rules. As can be seen in the following figure the university makes a contract with the publisher that gives persons with certain attributes such as staff members access to a number of eJournals. The researcher is contractually related to the university as staff member. When trying to access a paper the publisher will first ask the user to authenticate at the university so that some user attributes such as "is-staff-member" will be exchanged. Then the publisher will give access to the paper.



This dedicated federation is based on two contracts and the trust that the university handles user attributes with care. The mission is well-defined for both sides: the university wants to give researchers access to all relevant publications and the publisher wants to ensure his income. The additional rules required by this grid/federation are comparatively shallow, since they only have to make specifications about the service to be delivered to certain members of the university, its technical implementation and the trust in the universities' correct behavior. It may also make statements about the Quality of Service and specify penalties in case of misbehavior. This concrete model fits with the commercial model, however, in terms of our earlier discussion it is certainly a "shallow federation", since the number of additional rules will be small.

Summarizing, we can describe a number of characteristics that are typical for federations in the academic world:

- The partners share a **mission** that has to be made explicit and that every partner has to agree with.
- The partners have to describe the **trust relationship** which they all agree with, since in the strict sense their federations do normally not fall into the category "commercial model of grids".
- In general the partners in academic federations retain most of their **independence**, the federation just defines the regulations of the resource integration layer.

[9] The term "model of Grid Computing" is seen here as a synonym for a certain class of "federation models".

- The system of **regulations** is expected to be shallow, since topics such as quality of service are not an issue requiring severe penalties and since the ownership of resources will not be changed.
- **Penalty** regulations have to be defined in case of misuse, but since rights are not directly involved these can be kept simple. In general, exclusion from the federation will be sufficient which would require rules to decide this issue.
- Federations are not made for a short period, but they add facilities at a structural level that have to be maintained with a **long-term perspective** to satisfy the needs of the researchers.
- According to Volanis and Dumortier this type of federation falls under the "**Information Society Services**" legal framework at least within Europe.
- A set of **technological agreements** have to be accepted by all partners to get the federation operational. Processes have to be defined how to maintain these agreements over the years and how to adapt them to new requirements.

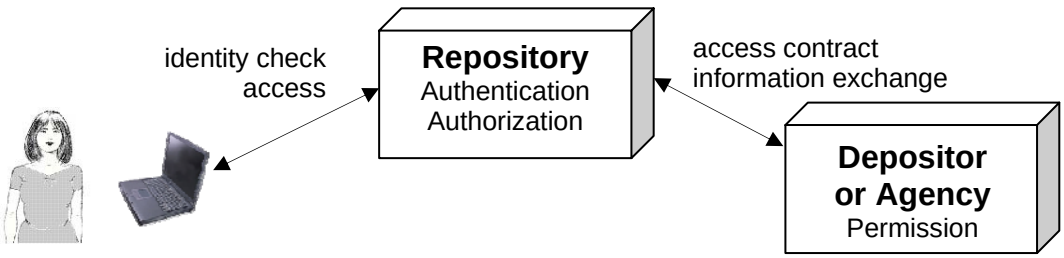
Federations in the academic domain turn out to be dynamic, i.e., new partners will join, others will stop their participation.

3. Rights Issue for the LR Domain

Essential to all regulational aspects in the language resource domain are issues that have to do with rights. To clarify the scope of the term "federation" we need some analyses of how a grid can influence the rights situation.

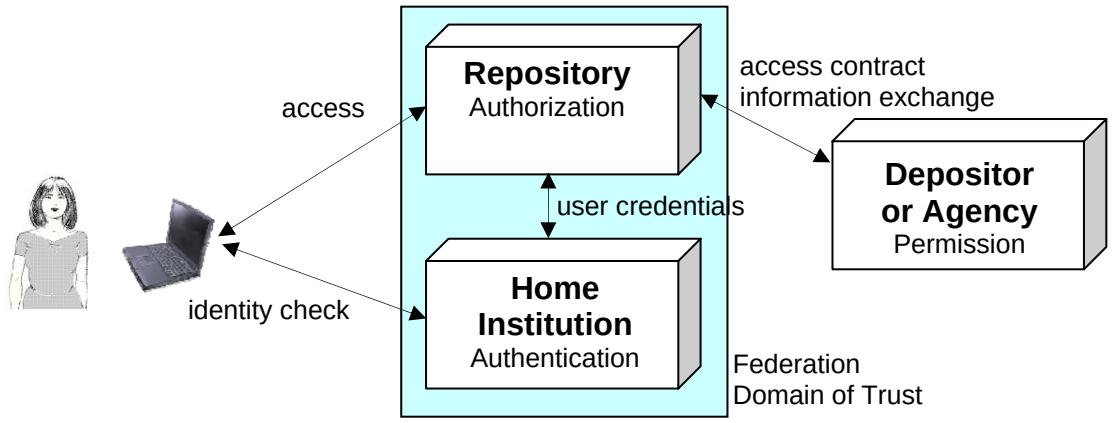
In general we have three important players when accessing language resources. We have the user who wants to access a certain resource that is stored in a repository. In general the resource is deposited by a researcher who has all rights in it[10] or it is provided by an agency that has these rights. In some cases the repository may have all rights in a given resource. Then the Repository also takes the role of the Depositor/Agency. In this chapter we will discuss a few scenarios where we will exclude the simple case that a resource is openly available via the web or where the resource is not accessible at all for anyone.

Scenario 1: This is the normal case where a user is dealing directly with the repository and where in some cases the repository will ask the rights holder whether access can be given. The repository takes full responsibility to handle access matters at a technical level as well.

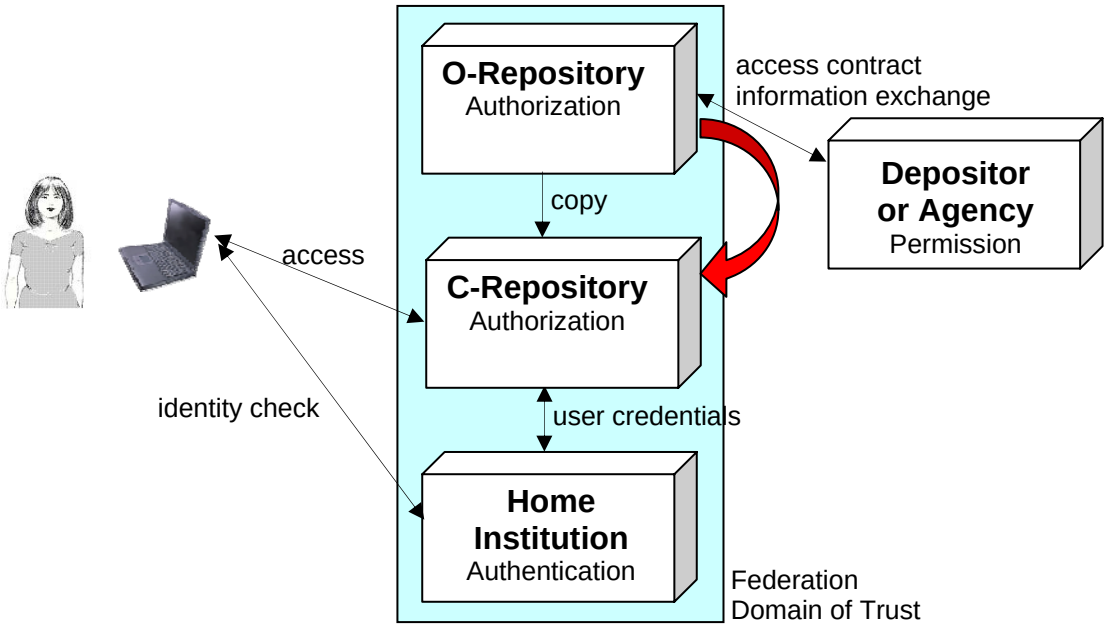


Scenario 2: In this scenario some additional components are introduced so that different instances form a "Shallow Federation". In the most simple case this just means that the functions "authentication" and "authorization" are split. A user that wants to access a resource has to first authenticate with his home institution which sends some agreed credentials to the repository, i.e., the repository relies on another instance to identify a user. The rights issues are not changed at all which makes federations of this sort very simple to establish. The trust relationship in the federation has to be specified, since we trust other archives to authenticate the users, so that they can be given access on the basis of trust.

[10] We assume here that the repository has the right of archiving the data.

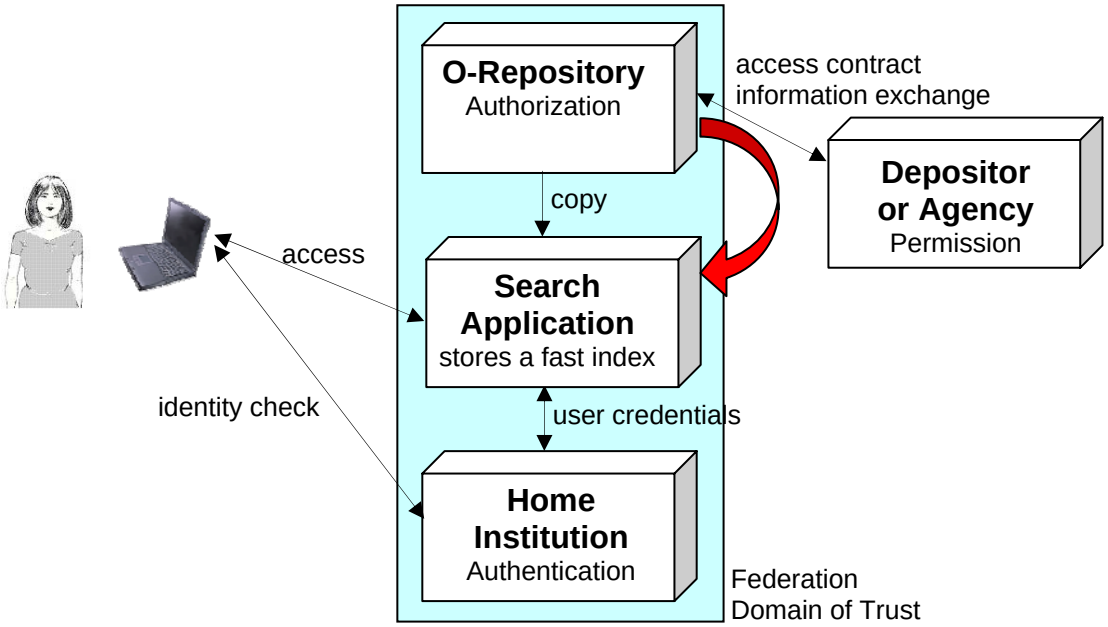


Scenario 3: In this scenario we assume that a resource is copied from the original repository to another instance which we call copy repository for several reasons such as long-term preservation and load distribution. This complicates the scenario slightly since the user does not interact anymore with the O-Repository that established the contracts with the depositor or agency, but with the C-Repository that does not have such a contract and probably even does not know any of the contract details.
The solution to solve this problem at a technical level is comparatively simple, since we only have to ensure that the rights on resources go with the copy and that the O-Repository (original copies) still is the only instance that may change them. Actually the technical solution would be different: the C-Repository would check at the O-Repository what the rights situation is and whether the requesting user is authorized to access a given resource. For this case the federation needs to be augmented with another trust relationship between the two repositories and probably some formal rules of behavior.



Scenario 4: We can think of a few mixed scenarios that can become very complicated to handle. These can emerge when for example applications are used that may be associated with graded access policies. As an example let's assume that a service provider wants to create an index about the contents of all resources in a number of repositories.
Of course, creating a fast index actually means copying the data and representing it in a different form that is optimal for searching processes for example. In the figure below one of the various possible architectures is shown where the service provider running the search engine will receive a copy of all data to create the fast index, i.e., all data is copied to serve a new type of application. In other architectures the service provider would just receive the query and will send it in a formalized form to

the O-Repository that has its own fast search engine operating on the content. This does not imply a copy of the data, but nevertheless searching means to access the contents.



Probably, this type of access was not part of the contract between the O-Repository and the Depositor/Agency. This could be solved by amending the contract, but such operations are very costly and difficult, in particular, since there will be other type of applications as well. More simple is to stick with the former rule that any access to the content has to be granted according to the rights of the user launching the query. Technically this can be implemented by checking the access permissions for any resource that is accessed in the index or that results in a hit[11]. At the management level this can create a heavy load if there are no efficient management tools. Whatever the solution is the O-Repository has to rely on the proper operation of the application which requires a more careful consideration of the trust relationship and probably more complex regulations.
In the distributed case where the search engine is operating on the data at the O-Repository the responsible developers can implement all checks and algorithms that are required given the contracts with the depositors and they need not to rely on proper software from third parties. However, they need to invest in own software development that may be not feasible.

Summarizing we can say that a federation configuration does not per se make the rights situation more complicated, but that it introduces the need of new trust relationships. New types of services, however, can lead to rather complex situations.

4. Open Access

It is in the natural interest of researchers to have access to all digital resources that are available. In particular the web with its new possibilities allows to dream from a domain of digital resources free of barriers for the researchers. According to J. Taylor "e-Science [5] is about global collaboration in key areas of science and the next generation of infrastructure that will enable it". The Cyber-Infrastructure NSF report of the Atkins Committee [6] advocates for open platforms and referred to a Grid as an infrastructure for open scientific research. For specific domains (electronic publications) the e-IRG roadmap [7] even urges public funding for the development of scientific software because current Intellectual Property Right solutions are not in the interest of science and the president of the MPG asks for new legal regulations that are not in complete opposition to current scientific usage scenarios enabled by modern communication methods and compliant to the framework of Open Access [8}.

[11] At the MPI one big index is generated covering all hosted resources. Including a resource in a query will only be given if the user has access rights for that resource. This seems to be a consequent and safe policy.

Data Grids are the kind of basic infrastructures currently being built up to create domains that integrate resources from different repositories, i.e., overcoming at least institutional boundaries to enable enhanced access and collaboration. However, there are still many obstacles to make resources openly available to researchers:

- There are and will be many resources that need to be protected due to privacy, religious and similar reasons, i.e., recorded persons don't want to be visible to the whole world.
- There are institutions that need to make some money to maintain their service, i.e., access needs to be controlled and some fee is required.
- The resources are partly donated from agencies that impose a restricted access policy and/or that want to get some money back.
- In "How open is e-Science" Paul David and colleagues [9] distinguish between e-Science and Open Science and discuss reasons for access restrictions that emerge from the research process itself.

Although many institutions fully support the Open Access initiative mainly as a counter movement to current trends of selling our cultural heritage to private institutions we need to realize that there are and will be many obstacles that will require access restrictions and sensitive access management policies. These are fundamental to our domain and any federation will need to address both in technical and organizational sense.

Grid systems are being established to make access management feasible in the kind of distributed scenarios we are working on. When designed correctly they will not influence the legal situation between owners, resource providers and users, but simply require additional trust relationships.

5. Implications for DAM-LR

DAM-LR is a pilot project at small scale with a limited time span to try out suitable technologies to integrate language resource archives and to better understand the legal and ethical aspects involved. It is not an infrastructure project, although the participating institutions expect to maintain the integration layer beyond the project time. But there is no legal or financial binding of doing so. Therefore, DAM-LR follows the typical social model where the partners see the benefits of the integration for the researchers, indigenous people and other users.

The same is true for the world wide network of archives called DELAMAN [10]. With the help of the joined geographic representation as can be found at the Language-Sites web-site [11] one can easily find resources that are stored in the participating archives. Resources that were recorded from languages that are spoken in the geographic neighborhood can be found in close geographic locations. The map below gives an indication of this geographic integration despite all institutional boundaries.

This integration is done purely on the metadata level, i.e., this information is openly available anyhow. Accessing the resources themselves will always invoke the access managing components at the various repositories which is compliant to scenario 1 or in future also to scenario 2.

The DAM-LR partners agreed that at first instance scenario 2 will be implemented in all respects, however, its architectural design was made such that scenario 3 can be realized as well.

Summarizing, we can say that DAM-LR just needs a shallow federation framework and that the partners need to agree on a number of technical aspects as being defined in the definitions documents.

This Google Earth presentation of Language-Sites points to material from various languages stored at different locations such as DOBES archive (triangular icon), AILLA archive (salamander icon) and MPI archive (Greek symbol). This is just the beginning, since more archives have indicated their interest to join.

Technical Agreements

- every partner will offer his metadata descriptions as harvestable XML files to allow creating a unified domain of language resources based on the IMDI infrastructure[12]
- every partner will certify its servers and services according to the accepted TERENA TACAR list
- every partner will setup a PKI system and sign its certificates with public keys
- the Handle System will be used to manage and resolve unique resource identifiers
- every partner is a Handle Authority and therefore has full authority to manage its own postfixes
- every partner will maintain its Handle System properly so that URIDs can be resolved
- MPI will setup a mirror service for another Handle Systems as a test, but will not modify any records
- access rights records will be associated with the URIDs in a unified format and managed only by the owning institution; these rights hold for all copies of resources
- every partner has to maintain and serve the agreed user attributes
- all authorization information for a certain resource is exclusively maintained by the originating institution - this right is not touched
- the access rights information is part of the URID database; it is up to every partner how this access rights information is maintained
- the choice for an authentication system is left to the partner institutions and a password identification mechanism is seen as sufficient
- Shibboleth is used to exchange user information, it is left to the partner institutions how they couple Shibboleth with their authentication solutions to extract the user information; however, the partners have to ensure that this interaction is operating correctly
- it is left to the partner institutes to decide about their resource manager, the component finally lending access to a resource[13]; it is the responsibility of the partners to solve the interaction between Shibboleth and the resource manager component

[12] The IMDI infrastructure is a result of EU funded projects such as ISLE and INTERA.
[13] In most cases the Apache Web Server will play the role as resource managers.

- DAM-LR (Shibboleth) differentiates between resource providers and identity providers, other institutes may be added as identity providers to allow their users to access resources from the partner archives, if they accept the rules and agreements
- every partner will provide mechanisms to request access permissions for a certain resource, it is left to the partners how they do access management

Federation Rules
With these technical agreements in mind and with the knowledge that only scenario2 will be implemented we can derive a few rules for the DAM-LR type of test federation. Since in this scenario authorization is left with the institution where is was also in scenario1, only shallow rules have to be defined.

- the partners need to declare that they will handle all user information with care and that the user attributes are defined correctly
- the partners need to declare that access management definitions are left exclusively to the institution "owning" the corresponding resource
- the partners need to declare that they will maintain the required technical infrastructure as described by the technical agreements

For DAM-LR it is not necessary to have a certified portal that covers more than what is official for the DAM-LR project; in particular, there is no need for a unified set of rules describing correct behavior with respect to the usage of resources, since all access management aspects are handled by the "owning" institutions in the same way as until now.

Since there are no structural funds reserved for the period after the end of the DAM-LR project, there is no legal binding for the partners, so we cannot expect a persistent infrastructure.

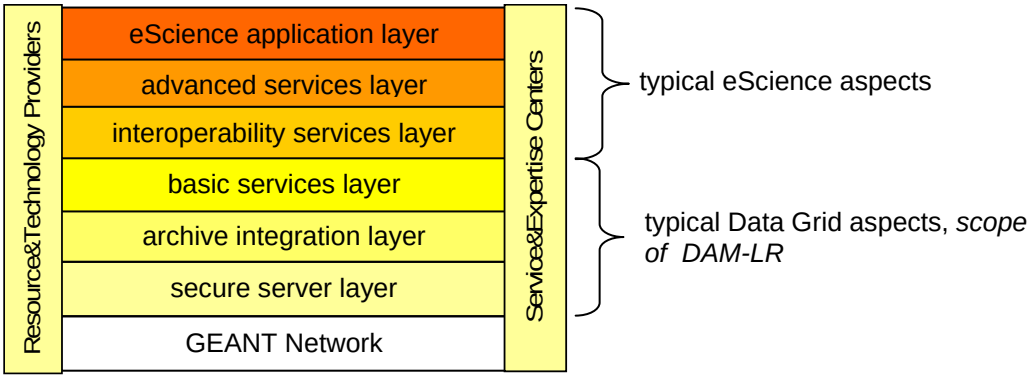
6. Grid Frameworks

Grid systems are to a large extent discipline-independent, therefore we see different Grid infrastructures emerging at different levels: discipline crossing at European level (EGEE), discipline specific at European level (DEISA, DAM-LR), discipline crossing at national level (D-Grid), discipline specific at national level (Text Grid). All these initiatives are important to get a deep understanding about the service layers to be provided, the components to be used, the differences between the disciplines and the responsibilities at national and European level. One of the big challenges will be to bring all expertise and know-how together and to come to widely agreed standards to achieve a high degree of integration and interoperability. At this moment one can speak about highly coordinated activities at the national level in particular in the UK and in Germany as far as we know. This is due to the long-term intentions that are driven by the ministries and research agencies. At the European level the close interaction has still to happen and the need for harmonization is enormous.

DAM-LR is a contribution from a specific data-oriented discipline in the humanities - the integration of language resource archives. Yet there are not so many experience gained in Grid projects in the humanities and social sciences (HSS) at a cross-national level. Even at national level there are not so many real Grid projects in the HSS. Therefore, the experiences gathered from DAM-LR are very important for further projects and research infrastructures at the European and at national level. Actually, the integration of repositories such as language resource archives only makes sense at a cross-national level, since only the virtual integration at cross-national level will create the new opportunities that researchers are looking for.

7. Implications for Research Infrastructures

Currently, the European Committee and the EU member states are discussing about building up persistent research infrastructures (RI) that will enable the e-Science scenario in various disciplines. Technically DAM-LR covers the basic integration layers of such a research infrastructure as indicated in the following figure.



It is obvious that RI have to go beyond what DAM-LR wants to achieve. Centers have to make formal commitments for a number of years with respect to the services they want to offer to the researcher community. They need to sign agreements where the even the Quality of Service (accessibility, availability) is specified, since otherwise the researchers can't rely on them. Here we can indicate a few such lower layer services that have to be guaranteed within a RI by a few centers for redundancy reasons:

- each federation needs to be augmented by official portals that contain certified and agreed documents
- several metadata portals need to be available to harvest metadata and exhibit a complete or split catalogue
- services to guarantee trusted servers and services are already maintained, i.e., all Grid projects can rely on them
- harvesting semantic mapping of metadata, metadata browsing and search services
- metadata schema registration service, metadata category registry maintenance
- Handel Services for institutes that can't run them themselves
- mirror services for URID resolving
- maintenance of formal lists (registries) of federation members and their characteristics
- it seems to make sense that a Europe-wide authority defines attributes for typical user types as they are accepted within the European research community, that all interested research institutions agree with the specifications and setup their user management so that these are maintained and can be extracted by authorized services; these specifications even should be synchronized at international level; they can widely be based on the specifications on the EduPerson agreements in the US
- certainly a European wide declaration will be needed that is signed by interested institutions that specifies that institutions will maintain user attributes correctly and that they will make them available to certified services in a federation; this general declaration will allow for example universities to join a merged resource domain without the need that for each Grid type of activity separate declarations have to be signed
- a list of accepted and certified components such as Shibboleth has to be maintained after extensive tests have been carried out; this guarantees that software components and applications that work with sensitive information such as user specifications are intensively tested before they can be applied in a federation
- of course many training services have to be offered

CLARIN Research Infrastructure

In the special case of the CLARIN infrastructure initiative [12] it will make sense to follow a two tier approach. On the one hand we should extend the DAM-LR Grid project to a pan-European dimension to build with some major goals in mind:

- extend all strategies to make them scalable,
- extend all mechanisms so that they can be turned over to a persistent research infrastructure,
- already establish the above mentioned services and give funds to the centers that will offer them,

- broaden the network of knowledge to the Europe-wide community, carry out training courses and setup a help facility
- interact with other Grid activities at European and national levels to achieve unification
- prepare the Grid layer so that it can be integrated into a persistent research infrastructure for language resources and technology

On the other hand it is necessary to explore and accommodate the requirements for the much more complicated and to a large extent discipline specific higher layers to be tackled in a research infrastructure. Further, CLARIN can be the umbrella to work out all financial, organizational and legal aspects of a research infrastructure.

References

[1] D. Nathan, R. van Veenendaal (2006). DAM-LR as a Language Archive Federation: strategies and prospects. LREC Workshop on Research Infrastructures, Genoa.
[2] http://en.wikipedia.org/wiki/Federation
[3] http://www.dspace.org/
[4] N. Volanis, J. Dumortier (2006). A European Legal Approach to Grid Computing. IEEE eScience Confererence, Amsterdam
[5] J. Taylor (2001). Presentation at e-Science Meeting by the Director of the Research Councils, Office of Science and Technology, UK, http://www.e-science.clrc.ac.uk
[6] D. Atkins, K. Droegmaier, S. Felman, et al (2003). Revolutionizing science and engineering through cyberinfrastructure. Technical Report, National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, D.C.: NSF
[7] M. Leenars (2005). e_infrastructures Roadmap: http://www.-e-irg.org/roadmap/eIRG-roadmap.pdf
[8] Open Access: http://www.soros.org/openaccess/
[9] P. David, M. den Besten, R. Schroeder (2006). How Open is e-Science? IEEE eScience Conference, Amsterdam
[10] http://www.delaman.org
[11] http://www.mpi.nl/services/mpi-archive/GE_language_sites
[12] http://www.clarin.eu