Appendix A: Impact of DAM-LR

Peter Wittenburg, Daan Broeder, Remco van Veenendaal, Sven Stromqvist, Vincent Wagelaar,
Thomas Schöntal
DAM-LR Project
29.12.2007

1. Introduction

DAM-LR was a comparatively small project and it was not planned as a project for policy building on a large level, but as a project to get hands-on experience in our discipline with introducing relevant technology components for federation building and with discussing the general requirements and implications of such technology when making real steps towards a real federation. For the partners the project had two essential foci:

- Becoming aware what a repository/archive infrastructure is and building a local one with all features that are necessary to enable an integration step.
- Based on these local repositories building an integration layer and study all aspects.

Therefore we see DAM-LR as complementary to initiatives such as TERENA, DRIVER etc.

Baseline
When the DAM-LR project was designed the consortium had no idea that there is/will come an ESFRI roadmap for European Research Infrastructures, that there will be a DRIVER project, that there will be an increasing number of national identity federations, which components will be seen as mature enough to be widely accepted etc. What we knew at that time was that

- repository/archiving issues became increasingly important due to the bad state of language resources in our field
- there was no broad awareness in the needs of metadata, unique and persistent identifiers, proper repository/archiving systems and specific service centers in our field
- only few people in the research domain saw the need of an integration layer for research data and technology for a number of different reasons
- there were just a few experts in Europe and almost no one in our field who had sufficiently deep knowledge about the technology and the implications of federations
- in the library domain a few pilot projects were started to establish national identity federations establishing trust relationships with some publishers
- infrastructure building is time consuming and that at that time there was hardly money to support such work

State
Without claiming too much we can say that compared to its small size DAM-LR had a lot of impacts at various levels - mostly of course in the linguistic domain. Having federation-ready repositories and an implemented integration layer we were invited to a number of strategic meetings, the Max Planck Society could not start building an organization wide AAI Infrastructure and without the DAM-LR experience, the CLARIN initiative would not have been started with the clear understanding of what we are aiming at.

In the following chapters we will describe a few special aspects where we claim that DAM-LR had influences.

2. Notion of Infrastructure

DAM-LR helped clarifying the notion of an infrastructure within and outside of the domain. The term "infrastructure" is relatively new in many scientific disciplines, in particular in the humanities. Only very few groups have practically addressed the aspect of infrastructure building that exceeds the narrow boundaries of a project or an institute. The resource and tools landscape is largely determined by isolated solutions with a huge fragmentation and all the inefficiencies in our domain as a result. As is noticed by many other experts already this fragmentation forms a severe obstacle on the way to an effective eScience scenario. The **ESFRI Roadmap document** states that we are living in the "***century of the complex systems***" to stress that simple models and answers will not help us to meet the big challenges that we will be faced with also in the humanities and social sciences to find answers on essential questions about the future of "minds" and "societies". Complex models will be based on data-oriented methods and require the efficient access to many different types of resources and tools and the easy combination of them.

Competitive research can only be done in future when this new eScience paradigm has become reality. Let us cite **J. Taylor** to document the close relation between eScience and infrastructures: "***eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it***". A new type of technological and organizational framework which is commonly called "research infrastructure" will be necessary to enable this eScience paradigm. Research Infrastructures (RI) are not just another set of short-term technology projects, but they are long-term investments in persistent and highly available services that researchers can rely on and that will help to overcome the fragmentation. RIs have a strong technological component and they require new and smart standards, but equally important are proper solutions for the organizational-administrative aspects. RI will also change the culture in which research is being carried out.

From various reasons the establishment of a research infrastructure was not yet a topic in our domain. DAM-LR initiated the discussion about RI not only in our field, but also in related humanities disciplines and helped clarifying what a RI exactly means. Due to the parallel preparations of the CLARIN proposal a European awareness could be achieved.

3. LRT Federation

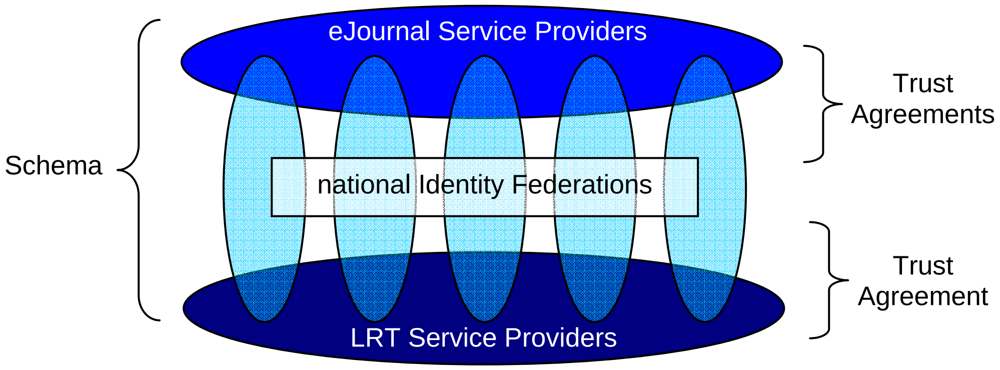
Yet it is still unclear how the final distributed landscape of services in the research domain will be shaped organizationally. Many initiatives are active at organizational, national and international level. But it will require hands-on experience and several competitive approaches to give answers to all the still open questions. It is obvious that federations will shape the ERA, but we don't know yet what the most appropriate structures will be.

We have seen in the recent years that a few national identity federations in Finland, Swiss, Norwegen and the UK were established and that other countries such a Germany, the Netherlands are in an advanced testing phase. All these federations want to establish trust relationships between them (representing a large number if not most of the national research institutes) and data providers such as the big publishers. The federation discussion including crucial points as the type of agreements and the usage of user credentials to be exchanged is dominated by the publishers' requirements - at least in Europe. DAM-LR is one of the few projects which we know of where these discussions were started with researchers as data providers and users in mind. From the intensive discussions with some of the identity federation experts it became obvious that the researchers wishes deviate considerably. Here we just want to refer to the usage of the "eduPersonEntitlement" attribute as one example.

More importantly it was DAM-LR that clarified our minds of the goal that initiatives such as CLARIN need to achieve to make an integrated and interoperable landscape possible. The group of language resource and technology providers needs to organize itself as a federation and workout requirements with respect to trust agreements, license agreements, business models, IPR statements etc. Yet we believe that there is a coherence of requirements in our domain that will allow us to come with a limited set of formal documents that we all share. If this will turn out to become true (which will be a topic in CLARIN) this federation can make contracts with the emerging national identity federations. This means that with one signature all researchers of a certain country would be a potential user of the offered services without having to deal everyone him/herself with each individual LRT provider as it is now the case. The LRT federation would act as a unified service provider as is indicated in the following figure.

As an example we can assume that a researcher wants to work on a virtual corpus of language resources for a certain short time period where the resources come from different repositories in different countries. It would be disastrous if such a researcher would have to negotiate access permissions etc with all resource providers, read all the different license agreements etc. The burocratic hurdles would be so high that the researcher would not start with such a work. This is the current state hampering to tackle many of the advanced research questions.

It is the goal of the CLARIN research infrastructure to follow this path, it is DAM-LR that helped clarifying these ideas.

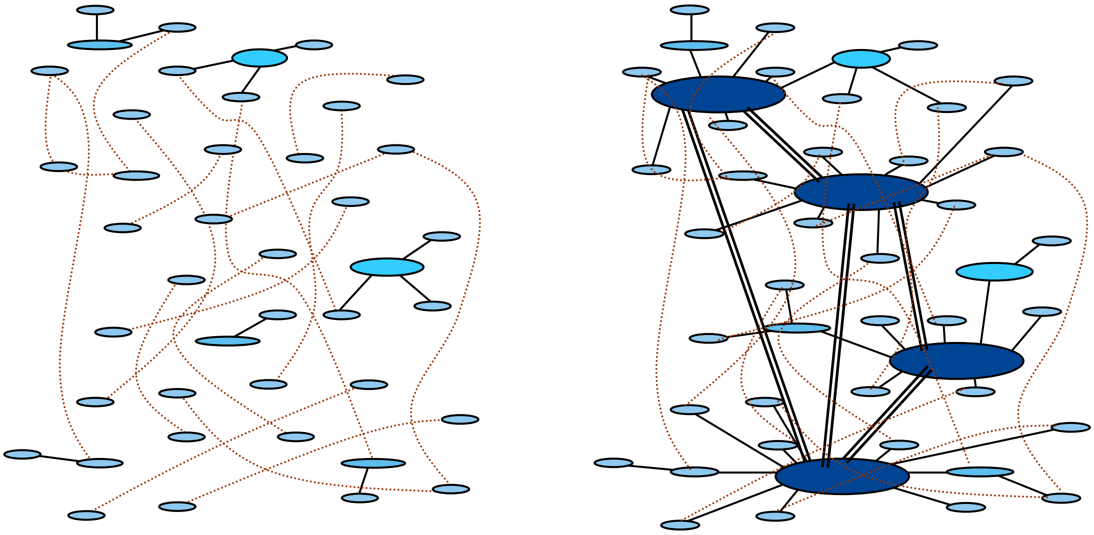


Any proposal to come to such a federation needs to be built on technologies that were tested out for our domain in DAM-LR. The experiences from the small-scale DAM-LR project are excellent, since they have shown us, what the requirements are to set up an AAI, what the state of the technology is and what the requirements for the participating centers are.

Based on the experiences from DAM-LR we also know that it is not at all obvious that every resource provider has the power to participate, i.e., only strong providers who have the internal capacity (management structure, repository/archiving system, expertise, etc) will be capable to participate. Again it is important for initiatives such as CLARIN to have gathered hands-on experience in this respect.

4. LRT Centers

Another important result of DAM-LR is that the notions of "center" and "repository/archive" system has become more obvious - at least for our domain.



Currently, the LRT domain can widely be characterized as an unorganized one in which small accidental networks and temporary collaborations exist. CLARIN intends to establish a structured domain of collaborating centers on top of these collaborations that are capable to offer stable and highly available services in a quickly changing world. The collaborations of researchers will further exist, since they are the key for all creative and innovative work, however, they will be released from all non primarily scientific tasks. This will only work when the individual researchers are not loaded with burocratic and administrative tasks and when the service centers are committed to offer simple to use services without burocratic obstacles. This integrated domain of service centers to be established offers users the possibility to work on even larger virtual collections with components gathered from various centers. The new landscape that for example CLARIN is aiming at is indicated in the figure shown above.

Such a landscape determined by service centers is the basis to overcome the fragmentation of resources and technology in the LRT domain. However, the infrastructure will only be accepted when the deal between researchers and centers creates benefits for the researchers:

- they hand over copies their resources and technology to service centers
- in return they will get seamless access without burocratic obstacles to a larger variety of LRT via stable and highly available services
- they will not be confronted with all sorts of license agreement details that hampers innovative work
- they can rely on a proper handling of access restrictions where this is required by privacy law or other serious matters

The "business model" has to support the typical work pattern of modern research not only in the LRT area, but in particular in the humanities and beyond:

- access patterns to resources are not predictable, they are dependent on the actual insights and questions which are changing dynamically
- often quick inspections on large virtual collections of resources originating from different resources are of highest relevance to see whether certain resources may contain answers to the questions in mind
- in the same way a tool box of tools is required to allow users to carry out a variety of operations on the selected resources in a simple way; users want to play and test, since often they are not sure whether a certain sequence of operations will lead to the answer they are looking for
- simple to establish workflows have to help researchers to overcome the interoperability problems, i.e., if a useful tool cannot be carried out immediately on a resource help should be given to transform the resource.

DAM-LR has shown that those institutes that want to participate as centers in such as federation need to invest funds to set the technology up and to understand and process all necessary agreements. CLARIN will be able to base its criteria on these experiences.

It became also obvious that each center needs to have a proper "repository/archiving[2]" system to integrate and serve its resources into a federation as different levels (metadata, PIDs, exchange of resources, portals). A number of such systems were studied such as

- **DSpace**: a digital repository for articles, books, courseware, journals, websites, theses and more; many institutions are using it already
- **FEDORA**: a system based on a clear object based container model, yet however with little functionality but with an active user group
- **SRB**: a very comprehensive system from the San Diego Super Computing centre in particular offering data distribution support which is used mainly in natural science environments
- **LAMUS**: a comprehensive system developed by the MPI for Psycholinguistics with a clear focus on language resources that is already deployed in a number of centers.
- **eSciDoc**: a new development of the Max-Planck-Society based on the FEDORA object model to make it useful for storing and accessing publication data primarily and later primary data
- **CMS**: a large number of content management systems are available in particular from companies

[2] Here we don't make a big difference. An archiving system needs to have a solution for long term preservation and accessibility of the resources.

Also in this respect CLARIN can draw upon the experiences from DAM-LR.

5. Live Archives

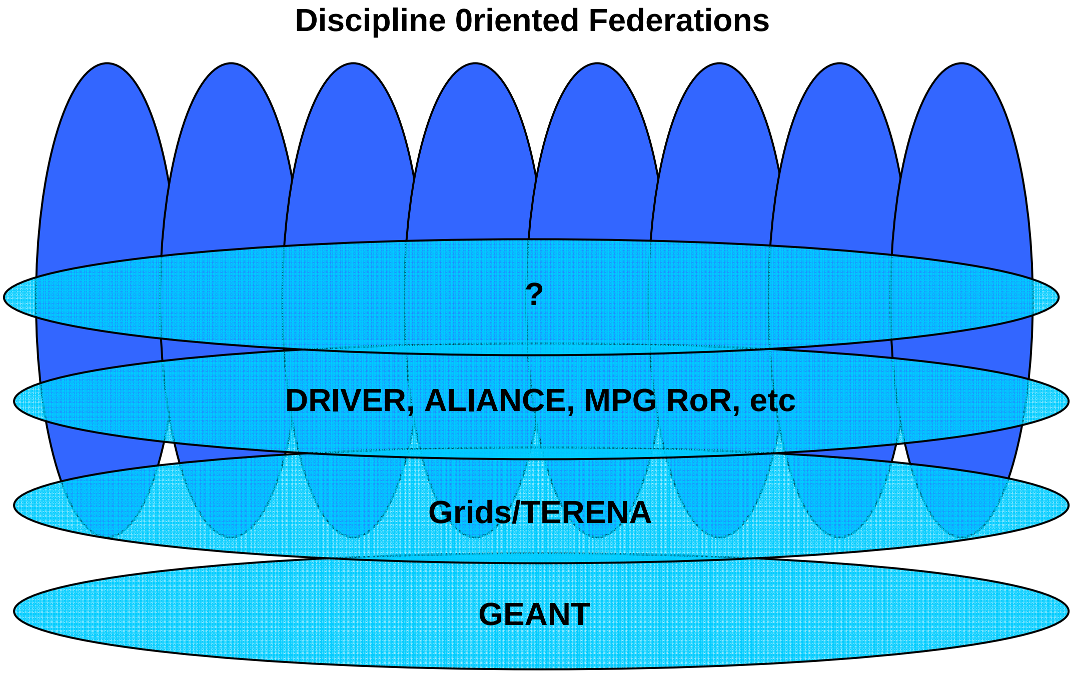
There is an ongoing debate what exactly the nature and task of a digital archive is. Some experts speak about a repository with a long-term preservation of static type of resources. Also the OAIS model provides this notion of an archive. In DAM-LR we worked out the notion of Live Archives that goes beyond the traditional models and defended this notion in several discussions. Digital archives are fundamentally "dynamic bodies" since we need to carry out migrations of media and formats to maintain accessibility and interpretability.

It is the possibility of digital storage that copies can be created without losing information[3], that we do not preserve the carrier and the form, but the content and that we know how to organize the resources in such way that existing ones are not modified by extensions or changes that allows us to build Live Archives. Live Archives offer access to all data and motivate users to extend and enrich the existing content without losing the authenticity of the data.

This discussion is not yet terminated.

6. Multiple Federations

As already indicated, currently there are many initiatives for setting up research infrastructures and federations. Some of them are discipline oriented with the goal to solve new challenges within the discipline and with an eye on cross-disciplinary services to related research areas. Such federations need to provide services that implement deep knowledge of the domain about primary and secondary resources, about the relevant workflows and the specific culture in the field. The discipline oriented approaches still reach out into the grid layer, since yet it is not obvious which grid-type services will be delivered in the different countries. This may change during the next years, still practical experience is required to establish clear support levels and interfaces.


Discipline 0riented Federations
?
DRIVER, ALIANCE, MPG RoR, etc
Grids/TERENA
GEANT

In contrast to these initiatives we see a number of cross-disciplinary initiatives at organizational, national and European[4] level. Grid projects are based on the excellent European network facilities and establish service layers such as compute grid infrastructures and national identity federations.

[3] We will not discuss the serious phenomenon of "concatenation" effects when using for example compression techniques.

[4] Some initiatives such as in genetics even work on an international level.

Yet it is not obvious which kind of services should be included. We expect for example that services offering unique and persistent identifiers at affordable prices[5] should be established at a national or even European level.

Some cross-disciplinary initiatives go beyond the grid level and want to establish cross-disciplinary federations. To be mentioned here as examples are the EC-funded DRIVER project, the ideas for an open ALLIANCE initiative in Europe, the EC-funded DARIAH initiative, the D-Grid project in Germany and the Registry-of-Registry project within the Max-Planck-Society. Traditionally these initiatives are centered around big libraries which makes sense since the type of resources (publications) exhibit discipline independence to a large extent and semantics of the metadata vocabulary used are also discipline independent to a large extent. The DRIVER project wants to make a step towards including research resources and the repositories that are storing them. Yet it is not obvious where exactly the boundaries will be, since they would easily leave the domain of their traditional experience. Even with respect to metadata one cannot speak anymore of a semantically coherent domain when including primary and secondary research resources. The disciplines use the metadata descriptions to facilitate research questions, i.e. the structures and in particular the semantics represent the domain knowledge. Creating a joint metadata domain would require complex semantic mapping technologies. Mapping all discipline vocabularies to the Dublin Core metadata set would severely reduce the semantics. The ALLIANCE initiative is even more vague in their goals and it does not seem that they exactly know what to do.

The DARIAH research infrastructure initiative has the goal to act as an integration layer for the humanities. Yet it is too early to make statements about the scope of the integration DARIAH can and wants to achieve. In the Max-Planck-Society there is the intention to join all metadata sets for society wide purposes, i.e. the goal is restricted. Nevertheless, the problem of semantic mapping will have to be solved.

It is left to initiatives such as CLARIN, CESSDA etc to closely cooperate with cross-disciplinary initiatives to find out where the optimal interfaces will be.

7. Awareness and Knowledge in the Field

DAM-LR is the first project in the humanities as far as we know that addressed the issues of federation technology and federations by gathering hands-on experience and not just by theoretical discussions. We have achieved the following aside from having a running installation:

- spread awareness in our domain by many presentations and papers
- at least in three centers we have now detailed knowledge about all technologies involved
- an excellent relation with one computer center in Germany and one in the Netherlands could be established that allows us to take profit from their deep knowledge in grid aspects
- we could train a few experts in our domain who can take further action

8. Outreach

In this chapter we want to highlight the major events where we could spread our messages and stimulated other communities to adopt the approach:

- for the EC-funded DRIVER project that wants to establish an interdisciplinary network of repositories beyond the libraries based on a metadata integration the DAM-LR partners are ready to participate
- some national identity federations have realized through the discussion with us that there is more than just the big publishers with whom they will make contracts in future
- the Dutch Grid Project has formed a collaboration with us with the goal to extend the integration to web applications
- the BOREAS project (within the International Polar Year) invited us to report on data management and distributed access

[5] The DOI services cannot be used to register all primary and secondary research resources due to pricing.

- the NSF invited us to a workshop on data management and archiving to present our approach and the perspectives
- in Brazil and Argentina we have stimulated the emergence of national centers for language and cultural resources in discussions and workshops that included high ranking ministry officials (these centers are eager to participate in European infrastructures)
- in the world wide DELAMAN network of archives we have presented the approach and it was widely accepted - yet Europe seemed to be the only place until now where a concrete project could be carried out
- by organizing the eHumanities workshops in the IEEE eScience conferences we could spread messages to other humanities disciplines on the one hand and could inform in particular the natural sciences that there are also serious activities in the humanities
- three of the first six eScience Seminars within the Max Planck Society were devoted to DAM-LR related topics
- the Max Planck Society decided to set up a authentication and authorization infrastructure by making use of the knowledge gathered in the DAM-LR project
-

It was already mentioned that the DAM-LR experience had a big influence on the construction and shaping of the CLARIN project and it can be seen from the presentation list that we were very active within our domain.

9. Political Statement

The intensive work during the last three years in the area of digital repositories, grid/federation technologies and digital libraries made obvious that the technology is dominated by US initiatives. Major components were developed by US initiatives and frequently they are based on excellent designs, i.e. all core knowledge is in the hands of US institutions. Despite all "open source" argumentation this dominance will continue to exist, if Europe does not apply a much more aggressive strategy and if it does not support European developments where they seem to be competitive. Partly it is just the smart marketing that brings our US colleagues an advantage although European have competitive technology. People too easily argue that "only initiatives from the big players in the US" will have a chance. This argumentation will perpetuate our dependency.