Deliverable 9.3/10.1
Distributed Solution Report
Adapt & Integration Report

DAM-LR

011841

Distributed Access Management
for
Language Resources

implemented as
Specific Support Action

Contract Number: *011841*

Project Coordinator: Peter Wittenburg

Project Web-Site: www.mpi.nl/dam-lr/

Deliverable: D9.3/D10.1

Authors: MPI

Responsible: MPI/SOAS

Date: 1.07.2007

Content

It is suggested to merge the deliverables 9.3 and 10.1 since in reality the subject matters could not very well be separated. Building and adapting was a continuous process as was already explained at the review meeting. We now entered the test phase where we also will need to carry out some remaining adaptations. The test scenario will be reported in deliverable D11.1.

Since there were only little changes with respect to the setup that was already presented at the review meeting, this report will be kept short.

1 Introduction

This is a report about the final setup with respect to the implementation of the various technical federation components in DAM-LR and about the federation discussion after having finished the construction phase and entering the test phase. We describe the configuration that was created within the DAM-LR federation and were appropriate highlight any local differences from the so called MPI reference implementation. The basis for this work are the specifications described in the D8.x versions and D9.2.

The metadata situation is described in the last D3.1 version. Here we only want to mention again that the repositories from MPI, INL, Lund and SOAS have a joint metadata domain and that IMDI is used as common platform. Currently, MPI is harvesting all relevant contributions and adding everything to its on-line browsable and searchable catalog.

2 PKI System

All partners were able to obtain the official EUGridPMA based PKI certificates as described in D9.1 using one is a prerequisite for transparent working of the Shibboleth middle-ware. These will be used in the tests.

3 Unique Identifiers

The Handle System (HS) from CNRI is used for the purpose of identifying archived resources. Since the initial DAM-LR agreement CNRI has implemented a new license policy requiring new applicants for a handle prefix to pay a small annual sum. This did not change the acceptance of the HS, as all DAM-LR partners accept this change in policy.

The HS system has currently been integrated into the reference implementation developed at the MPI in such a way that:

1. A Handle Server (for handle to URL handle resolving) was setup.
2. New ingested resources are associated with handles.
3. A tool is available that allows archive managers to move or rename resources and that will update the HS database with the new URLs.
4. Handles can be used to retrieve resources from the archive.
5. The archive catalog allows users to obtain the handles for the stored resources.
6. Handles are used in a language resource citation format that is being further developed

It was agreed earlier to use the HS to exchange authorization records. This would become a necessity when DAM-LR partners store copies of each others resources. Although this is something not foreseen as a result of the DAM-LR project, it is something that will be one of the following logical steps in a federation. At the technical-meeting in November 2006 a preliminary format for disseminating authorization records was agreed upon.

The state for the other partners is:

- Lund uses the reference implementation as described above.
- INL has obtained a handle prefix, has set-up a handle server and issued handles for resources. The handles are available through the INL catalogue

Further we can state that:

- All partners defined a postfix syntax defined in D8.1
- MPI is currently testing a mirror handle server. setup and will eventually mirror the handle servers of the other partners. These may decide to do the same.

4 Authentication

Authentication involves two groups of applications or functions. The first group is web applications that are used to manage and access resources. The second is when accessing a resource directly via its URL.

As described in D9.1/2 authentication components should be integrated with the Shibboleth middleware so the task of authentication will be performed by the home institute of a user when he

accesses resources or applications from other DAM-LR partners. The use of Shibboleth also means that authorization or at least the authorization for required transfer of user attributes becomes mixed with the authentication issue.

As explained in D 9.1/2 OpenLDAP was chosen as an authentication component for the reference implementation. The OpenLDAP server has been installed and is operational in the test environment of the MPI. An LDAP schema was implemented according to the specifications describing in D9.1. It is clear that because of the sensitive nature of authentication and authorization we only can implement these components in our production environment if we have passed extensive testing. It was recognized that within the DAM-LR federation we would need a unique user identifier space to administer each other. A simple way to achieve this is to use an organizational id as prefix to each users (local) identification separated by a semi-colon.

Further testing has taken place and shown the authorization using the federation wide unique user identifier to work. However current contacts with other identity federations may lead us to reconsider the current implementation of the unique identifier format.

The synchronization of the MPI's own local user records, currently stored within Active Directory Service, with the LDAP is performed by a MBean application, that regularly copies the records of the MPI's local users to the LDAP. The list of copied attributes and their new names can be easily changed while the synchronization period is a configurable parameter.

The archive web-applications of the reference implementation will authenticate and obtain user attributes using an API to access the LDAP. The API has been implemented and allows authentication but does not yet offer access to the complete set of user attributes. The current local setup of authentication with MS Active Directory Service for the MPI local users precludes simple copying of passwords to the archive user records LDAP. Special programming logic takes care that for authenticating users without a password present in the LDAP, the authentication is done via Kerberos with the ADS.

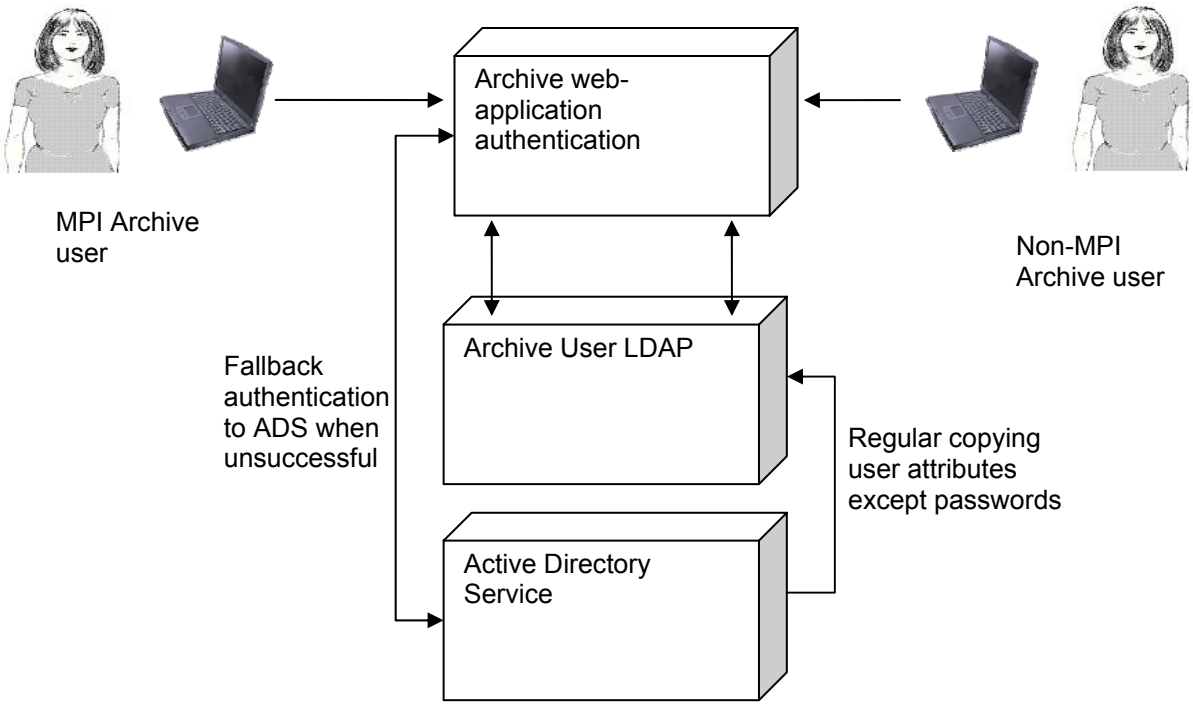
When accessing resources directly via their URL, the authentication and authorization mechanisms of the Apache HTTP web-server are used[1]. Shibboleth offers a SP (service provider) component that is integrated with the Apache web-server and that will redirect users to authenticate with their home organization. The redirection mechanism was already shown to work (tested between MPI and INL) end of 2006. Currently the user attribute exchange has also been implemented and was shown to work between MPI and INL and MPI and Lund. The set of exchangeable user attributes for the DAM-LR federation was established already some time ago. However new insights and the desire to conform to existing practices and allow easy future joining of other identity federations may yet prompt us to make some small modifications to become completely mainstream (see below).

The Shibboleth IP (Identity Provider) component is where authentication requests from the SP are processed. In the MPI setup we need to be able to create the same fall back mechanism for local MPI users as when authenticating for web-application as described above. This is arranged by making the IP use the security mechanisms of a Tomcat Servlet container that is configured to use a JAAS realm. The JAAS realm allows falling back to ADS authentication if authentication with the LDAP fails, just as is the case with the authentication API.

The state of affairs with respect to the DAM-LR partners

- For testing the IP and SP providers were installed, and shown to be working on a test archive, the installations will be deployed on the actual MPI archive. This has to be carefully executed because of the very sensitive situation of the MPI's production environment.
- Lund is using a copy of the MPI reference setup except that they connect to a LDAP user of their own design.
- INL has a working IP and SP that covers large part of their archive..

[1] For media streams produced by the Darwin Streaming Server and accessed via the RTSP protocol there is no protection. This problem was bypassed by creating temporary links.



5 Authorization

The Shibboleth SP that is connected with the Apache web-server supports the use of authorization records in the same fashion as the normal Apache web-server, which in the case of the reference implementation is storage in the htaccess file(s). However the operation within the DAM-LR federation requires that some extra issues need to be addressed:

- In the authorization records we need to use the federation wide unique user identifiers.
- The IP needs to provide as the user identifier the federation wide unique one. This can be achieved by configuring the IP such that it creates the user identifier attribute by concatenating the appropriate user attributes from the LDAP with an archive identifier.
- In the reference implementation authorization records are created by an existing software component: the Access Management System (AMS) that must be able to draw on an existing list of potential users. However non-local users should be added to this list when necessary but only when their credentials are trustworthy. This can be achieved by a function of a new component, the Resource Request Service,(RRS) where a user can specify his particulars and request access to specific resources. The RRS can check the user's credentials by matching the information given by the remote Shibboleth IP. Currently the RRS is now operational although its connection to the Shibboleth IDP needs yet to be implemented.

The integration of the RRS will be tested in the test phase.

6 Federation Aspects

After having discussed the first note about federation issues intensively at various occasions (RI Workshop at LREC, DELAMAN workshop, DAM-LR workshops), a new paper has been created that relates the set of rules needed for a federation with the architectural solution chosen. It was already presented during the review meeting.

In the meantime several intensive discussions have taken place with existing identity federations such as the HAKA federation in Finland, the SWITCH federation in Swiss, the AAR federation test bed in Germany and a test bed in the Netherlands. The Athens federation in UK is organized in a different way, but will later also step over to a Shibboleth federation. All these federations were mainly focusing until now to allow researchers to access electronic publications (Elsevier, Ovid, etc) and other commercially available resources. From these federations it is known that the exchanged attributes and attribute values are based on eduPerson and InetOrgPerson. To allow shallow management

about 6 attributes are mostly seen as being required and there is an ongoing debate whether publishers require the eduPersonEntitlement attribute to be filled in although this would require lots of overhead.

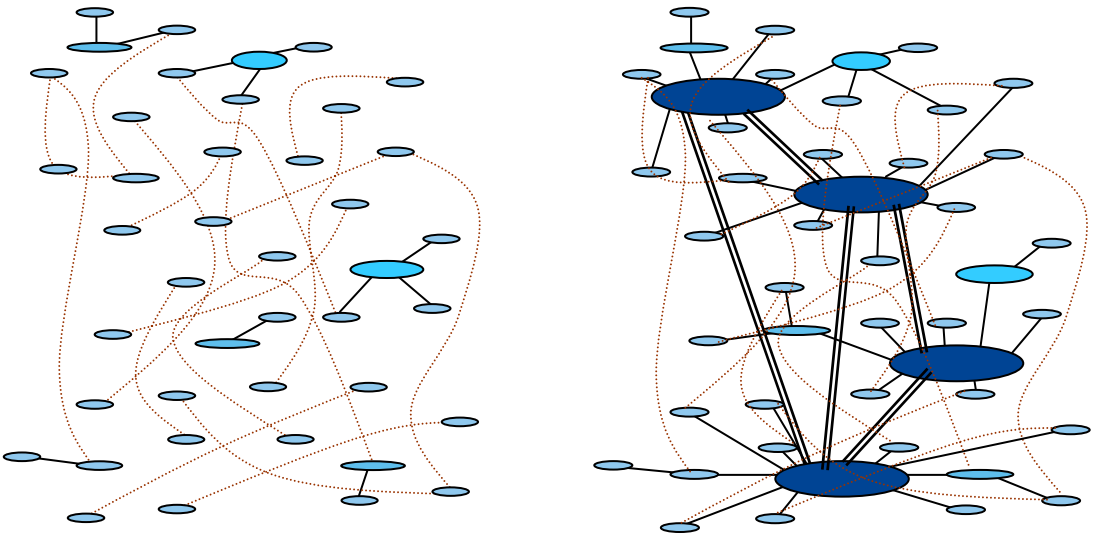
This paper, however, is not the place to present the state of the discussion. Important is that with DAM-LR for the first time as we know humanities archives sat together to discuss from their perspective as scientific resource providers what there requirements would be. This will allow us to influence the discussions about the kind of attributes and values to be used. What we can see clearly that these requirements are not the same as for some of the publishers. DAM-LR will, therefore, document the requirements and will provide them to initiatives such as CLARIN for further elaboration and TERENA to agree upon a widely used vocabulary. It is obvious that one needs to come to suitable compromises at the European level that fit the researcher's needs as well. Since DAM-LR is a small project, we will rely on the discussion process in CLARIN to make statements at the European level.

7 First Experiences

After one year of hard implementation work much experience has been gained which will be transmitted to the CLARIN work. We can clearly state that

- installing and integrating all components (joint metadata domain, joint URID domain, joint user domain) in a distributed fashion is something that requires strong institutions, since the overhead is considerable and the required level of detail IT knowledge is high
- in particular the proper setup of the Shibboleth component is still a complex affair requiring specialist knowledge
- there are two aspects that require much attention: (a) the integration of Shibboleth with the local authentication environment and (b) the integration of Shibboleth with the local authorization environment. In most cases some workarounds have to be applied to get the different components smoothly cooperate

The message for a CLARIN like research infrastructure is therefore that there are probably not too many national centers that can participate in a full-fledged scenario as worked out by DAM-LR. The costs would be too high for the construction as well as for the operation phase. At the ECRI meeting in Hamburg Peter Wittenburg described the change with the following words: "It is widely accepted that eScience enabling research infrastructures will replace the current domain that can widely be characterized by accidental networks and temporary collaborations by a structured domain characterized by collaborating centers offering different kinds of stable and persistent services. This backbone of collaborating centers will be islands of robustness and availability in a quickly changing world.



The temporary collaborations of researchers will continue to exist, but they will be released from all primarily non-scientific tasks. It is the task of the centers, however, to avoid that the researcher's need

to carry out data-driven research will be hampered by the traditional “centralist” imago of these "centers". A new style of service-orientation will be required to implement an infrastructure enabling eScience.

8 Testing Phase

Our primary objective is currently to test the Shibboleth-based configuration and still adapt it where necessary:

1. Introduce the tested Shibboleth solution in the production environments if this was not already accomplished.
2. See if the current agreements about the exchangeable DAM-LR user attributes are still appropriate.

The Resource Request System (RRS) in the reference implementation at the MPI will be connected to Shibboleth, allowing users from the other DAM-LR partners to post access requests.

Currently there is work going on to establish the secondary handle services at MPI for the whole DAM-LR federation.

The intention to create shared metadata domain was realized but we will keep improving our metadata harvesting procedures and investigate if more efficient are possible e.g. regular incremental harvesting.

If time permits we could invest some time for the following subjects that could also be subject of the coming CLARIN project:

1. Investigate and create an efficient way for copying resources or synchronizing whole corpora between archives. Here we should address complex issues such as administrating ownership, access authorization (of which the theoretical planning is part of DAM-LR) and administrating the URIDs for copied resources.
2. Make the shared metadata domain more efficient by making the different metadata catalogs housed by partners that copy the MPI reference model “complementary” rather than “duplicated”. This would entail being able to parallelize metadata search queries over different distributed catalogs and being able to configure complementary domains of authority for a metadata catalog, although overlap is not harmful.

Further, as suggested during the review meeting we will improve the web-site and integrate a demo to show the functioning of the distributed domain. It is intended to organize a workshop to disseminate the results in particular to the language resource and technology community.