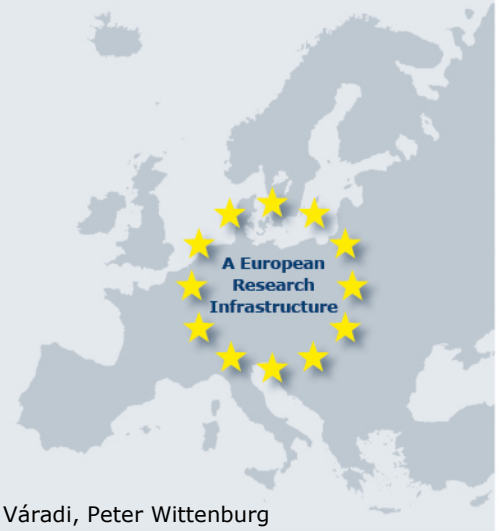
Common Language Resources and Technology Infrastructure

CLARIN

Language Resource and Technology Federation

2008-07-30 Internal Version: 6



Editors: Pekka Järveläinen, Sigfrid Lundberg, Jānis Džeriņš, Tamás Váradi, Peter Wittenburg

www.clarin.eu

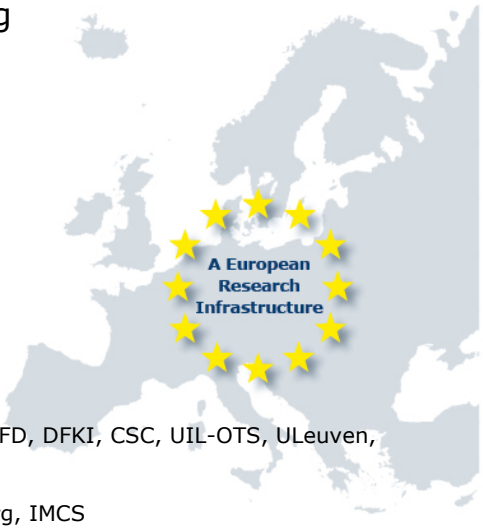
Language Resource and Technology Federation

CLARIN-4/2008
hdl: ??????

EC FP7 project no. 212230

Deliverable: D2.2 - Deadline: 1.7.2008 (postponed to 1.10.2008 due to late start)

Responsible: Peter Wittenburg



Contributing Partners: MPI, INL, OTA, RACAI, WROCUT, UPF, ELDA, ILSP, ILC, USFD, DFKI, CSC, UIL-OTS, ULeuven, ATILF, UTuebingen, HASRIL, CST, UTartu

Contributing Members: ULeipzig, CELTA, TILDE, Meertens, IDS, DANS, SBGöteborg, IMCS

Scope of the Document

This document describes the requirements for the Language Resource and Technology Federation that CLARIN wants to build up based on a stable network of centers as described in CLARIN-1-2008 and CLARIN-3-2008. It is also referring to a detailed discussion of possible solutions for persistent identifiers as described in CLARIN-2/2008.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

In chapter 1 it is explained why federation technology is an issue for a research infrastructure as CLARIN. In chapter 2 we will discuss various models of federations, distinguish identity and service provider federations and describe a few pillars federations need to have. In chapter 3 we will describe the technologies required to implement a CLARIN federation and in chapter 4 the major middleware components are introduced to establish a distributed authentication and authorization domain. In chapter 5 we summarize the requirements relevant for CLARIN and in chapter 6 we outline the procedural approach.

CLARIN References

• CLARIN Centers Types	CLARIN-1/2008	May 2008
• CLARIN Centers	CLARIN-3/2008	August 2008
• CLARIN Persistent Identifiers	CLARIN-2/2008	May 2008

Contents

1. Introduction

In the documents about CLARIN centers we described that dedicated and well-funded language resource and technology centers will be the backbone of a federation of service providers and already a number of criteria for the different center types were established. In this document we want to describe in more detail what is meant with the term "LRT Federation", what kind of requirements can be drawn for CLARIN and how we can come to a technical implementation of such a federation.

The term "federation" received considerable attention in the IT domain. The underlying reason is that there are an extremely increasing amount of web applications that emerged independently of each other and each of them having a separate user administration. The result is that users have an increasing amount of different identities prohibiting easy crosswalks by implementing for example a single sign on principle. It is obvious that distributed research infrastructures need to overcome this fragmentation. This is very well known to the IT community and the increasing number of national Identity Federations and the attempts at European level to harmonize between them are clear signals that the experts are looking for possibilities to overcome this bottleneck.

Although IT experts are working on these issues they are nevertheless a topic to be addressed by CLARIN:

- We can state that the knowledge at the institutions that can be potential CLARIN centres is very limited. CLARIN needs to bring together various experts from the LRT domain with those from the "federation" and grid communities.
- The experts working on "federation" technologies are largely driven by the requirements from the big publishers. Yet Service Provider Federations such as CLARIN were not identified as partners that will come along with own requirements. Yet there are no widely agreed rules according to which such community driven federations need to organize themselves.
- The experts in harmonization between the different federations that are emerging are not yet driven by community requirements.
- Establishing a federation is not just limited to creating a unified authentication and authorization infrastructure. Other pillars such as a joint domain of persistent identifiers that can be resolved, a joint security domain and a joint metadata domain need to be addressed as well. Some of the underlying problems to be solved are domain specific others not. Even in the case of those pillars that are not domain specific we cannot yet rely on satisfyingly running systems. Therefore, CLARIN needs to tackle them and perhaps offer services to other infrastructures.

This document is meant to describe all important pillars of the CLARIN federation and to derive the requirements with the intention to formulate the basis of the actual work to be carried out in the preparatory phase.

2. Federations

The term "federation" was introduced to describe the need to establish trust based on formal agreements when working in distributed networks allowing resource access to users who are accepted actors in such networks based on their affiliation and role. Before turning to federation technologies we first will discuss terminological issues, scenarios, the current activities in the area of identity federations, the nature of the proposed LRT provider federation and the implications for CLARIN.

2.1 Terminology

When consulting Wikipedia for the term "federation" [Federation] 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. We can speak 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 worldwide integration of open web content. All participants share the same common belief in the usefulness of worldwide data mining, i.e., they share the same mission.

In the [DSpace] domain users of the software discuss turning the user group into a federation that shares a number of interests that require a minimal governance rules such as enabling communication within DSpace community, ensuring the community is healthy and resolving conflicts of all sort. Also in this case we can speak about a shared mission and a very loose definition of membership.

N. Volanis and J. Dumortier [Volanis 2006] distinguish between two models of Grid computing[1] 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 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.

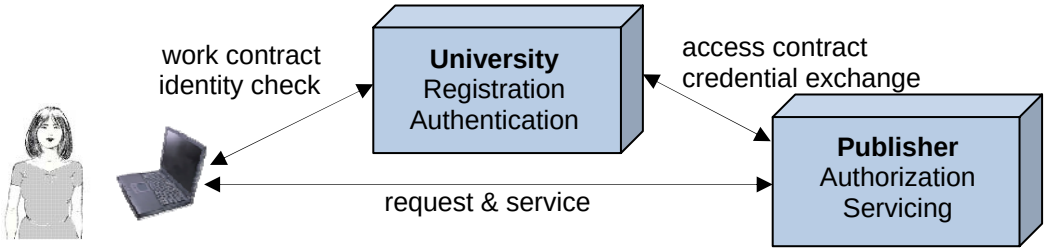
On the other hand, the **commercial model** sees in grid computing various business exploitation opportunities, i.e., companies need to control the resources to guarantee a Quality of Service. A number of enterprises can also form a Virtual Organization to share their data and resources based on a contractual 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 member 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 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 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

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

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 not normally 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.
- The partners have to agree about a system for classifying the user roles since they can play an important part in authorizing users for specific types of resources.
- In general the system of **regulations** is expected to be shallower, since topics such as quality of service are not an issue requiring severe penalties and since the ownership of resources will not be changed.
- Audits verifying the correctness of a partner's assertions about its users, may have to be allowed if the federation also has commercial contracts e.g. with the publishers in the above example.
- **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 just 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.
- Disclaimer statements need to make clear that no liability for malfunction or bad quality of service is taken.
- Since some institutions are bound to generate income for their resources and tools accounting mechanisms need to be integrated which is orthogonal to federation concepts such as single sign-on etc.

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

2.2 Scenarios

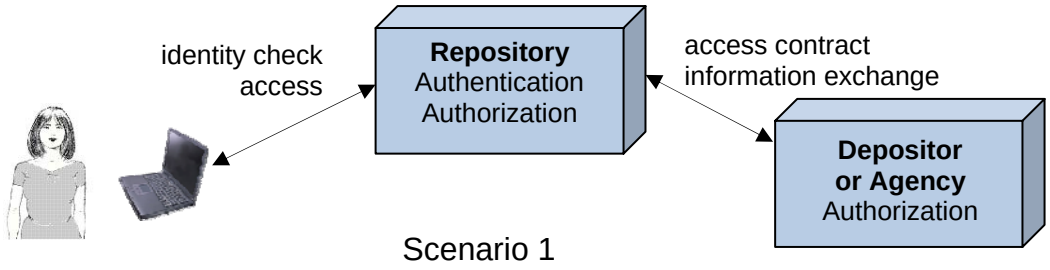
Rights issues are central to all federations in the research area, if all data would be open access we would not need to establish these domains of trust. To clarify the scope of the term "federation" we need some analyses of how a grid can influence the rights situation.

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 on it[2] or it is provided by an agency that has these rights. In some cases the

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

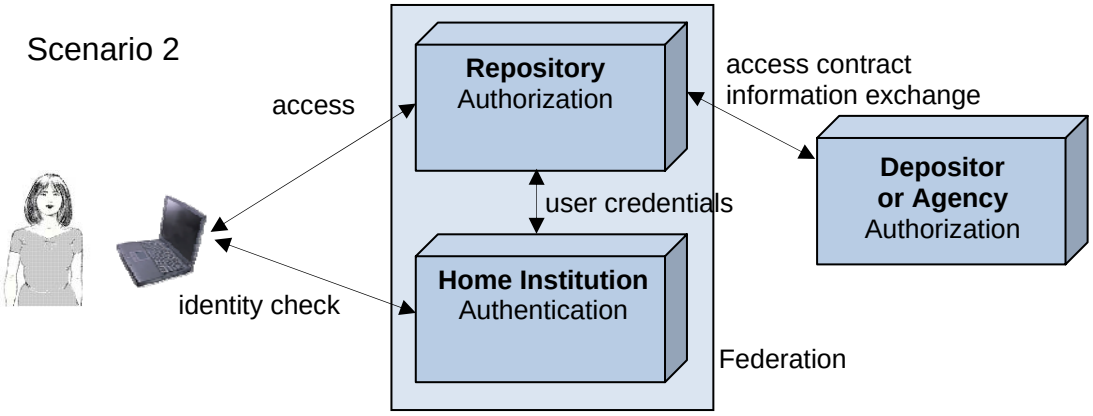
repository may have all rights on a given resource. If so, then the Repository also takes the role of the Depositor/Agency. In the following we will discuss a few scenarios where we will exclude the simple case that a resource is openly available via the web or where the resource is not accessible at 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.



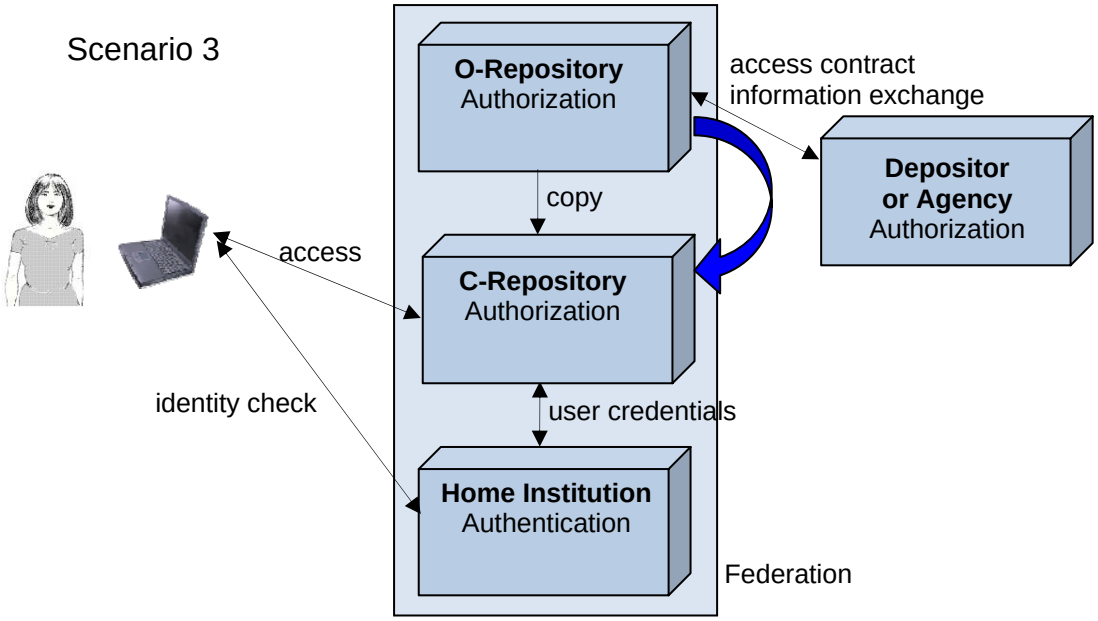
Scenario 1

Scenario 2: In this scenario some additional components are introduced so that different instances form a "Shallow Federation". In the simplest case this just means that the functions "authentication" and "authorization" are split. A user who 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 right users, and give them access on the basis of this trust.

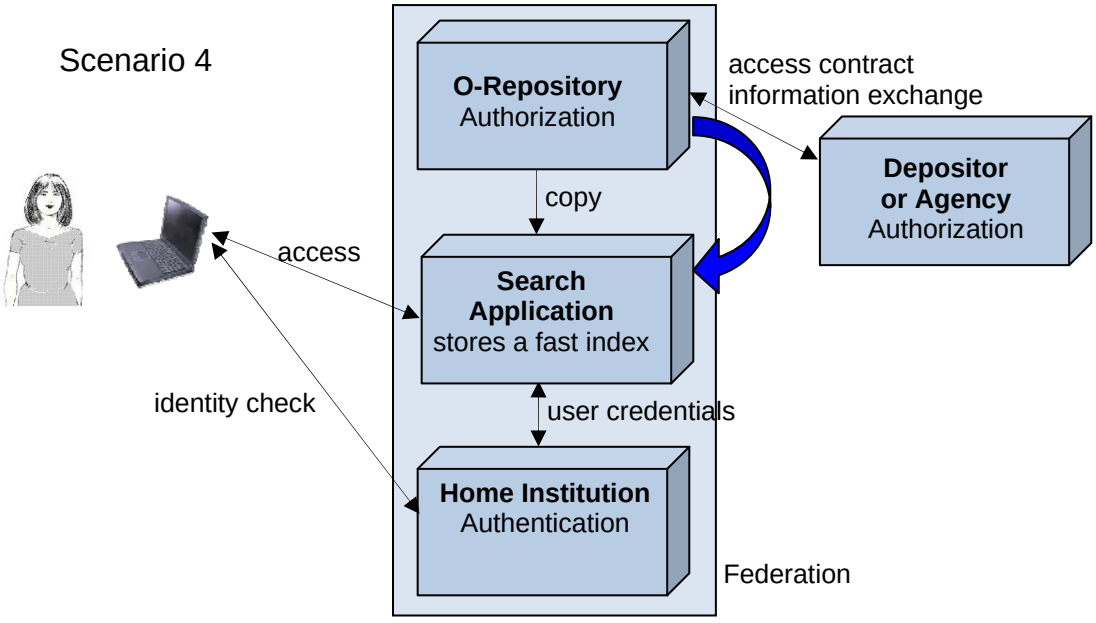


Scenario 3: In this scenario we assume that a resource is copied from the original repository to another instance which we call copy repository for several reasons such as long-term preservation and load distribution. This complicates the scenario slightly since the user does not interact anymore with 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[3].



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 different possible architectures is given 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 (using the SRU format [SRU] would be possible) 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.

[3] It should be noted that access to single words and sentences from a corpus is legal as long as the document cannot be reconstructed as a whole, since there is no copyright on words and sentences.

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[4]. 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.

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 is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" [Taylor 2001]. The Cyber-Infrastructure NSF report of the Atkins Committee [Atkins 2003] 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 [e-IRG] even urges public funding for 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 [BOAI].

In reality, 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 by 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 [David 2006] distinguish between e-Science and Open Science and discuss reasons for access restrictions that emerge from the research process itself.

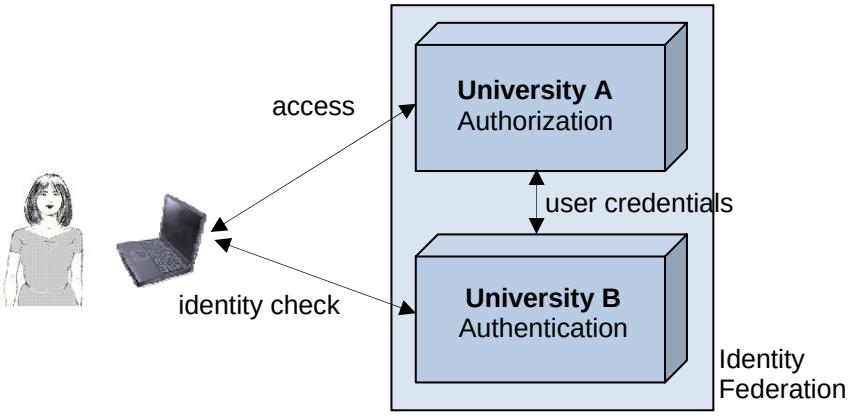
Although many institutions fully support the Open Access initiative mainly as a counter movement to current trends of selling our cultural heritage to private institutions we need to realize that there are and will be many obstacles to Open Access. These issues will require the implementation of access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

Data 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.

[4] 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.

2.3 Identity Federations

Increasingly more national research agencies understand the potential of identity federations where research organizations accept each others assertions about their employees. These national IDFs define a set of rules about user management and user attributes to make official statements about their employees, their role, the contractual state etc. In doing so they are creating a domain of trust where everyone can rely on the assertions about users. This principle of trusting each other was already used in allowing student classes from one university accessing online teaching material from another university. A student is registered at University A where he is also registered for a certain class. When accessing the appropriate teaching material at University B the two institutions simply exchange this part of information that the student, identified as such by his home university, is allowed to access the corresponding material. The two universities simply need to exchange one note that states that the specific student class is allowed to access this material. This can be very efficient. However in the humanities domain, researchers often work as individuals or in various temporary virtual research groups, i.e. the access management will in general not be as simple as in the case of a group of students having one (temporary) role.



Another benefit of such identity federations is that they offer the possibility of single sign-on. Assume that there are multiple institutions that have data the user wants to use - perhaps even in parallel. Technical arrangements can be made such that the user only needs to authenticate once and that the various components can recognize such an established authenticated session and are able to exchange other user credentials transparent to the user. One login would thus be sufficient to access all kinds of resources from different service providers. Thus the user could build virtual collections and carry out operations on them without noticing the institutional boundaries. These virtual collections create their own requirements. For instance each is a new instance associated with a proper metadata description which will contain the necessary information to give acknowledgements to the creators for example.

Also other "service providers" such as publishers and software vendors have recognized the potential of this technique to simplify access management compared to the traditional methods of using proxy services or fixed IP addresses that create so many problems. For all institutions that are part of such an Identity Federation a single contract could be made with a publisher specifying that "all researchers of all members will have access to the electronic publications under certain terms". Several IDFs started making agreements on this basis with for example Elsevier, JStore, Microsoft and many others[5]. An actual list of European IDFs or corresponding initiatives can be found under TERENA/TACAR [TACAR].

The differences between the various IDFs are along five dimensions:

- the specific rules that guide the IDF
- the set of user attributes used
- the possible values for these attributes
- the technology that is used as middleware
- the way auditing is done

[5] Detailed lists can be viewed at the Web-Sites of the existing national IDFs.

These differences motivated [TERENA] to put efforts in harmonization in particular about the attributes and their values - both together being often called the "schema". This work resulted in the [SCHAC] schema that can play a very important role as a reference schema to facilitate interoperability at European level. Mostly the parameter sets are based on the [EduPerson] list that has been worked out in the US. But the vocabulary does not always match with what is necessary in European institutions. For certain reasons for example the Max Planck Society will need a distinction between directors, staff members and PhDs although all belong to the category "scientific staff". A complicating factor is that some publishers have special wishes with respect to the usage of certain attributes. Also the Nordic countries are busy with harmonizing between their national IDFs. For this reason they founded the Kalmar Union [KALMAR] to work on a Nordic cross-federation.

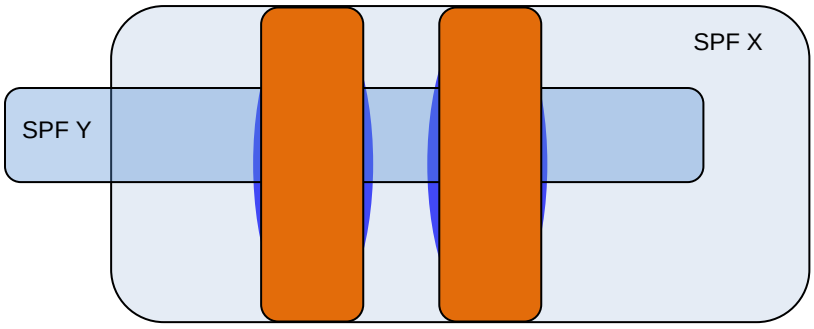
A working group in [GÉANT2] is running two projects. [eduRoam] for interconnecting wireless networks across Europe, and [eduGAIN] for interconnecting authentication and authorization infrastructures from the participating countries. There is progress to bring the two platforms together in particular, since the eduRoam protocol is dependent on exchanging passwords which is not acceptable for security and personality reasons. The purpose of eduGAIN is to provide the means for achieving interoperation between different Authentication and Authorization Infrastructures (AAI). eduGAIN wants to provide the technology necessary for carrying out these steps and thus interconnecting different AAI systems. eduGAIN is speaking about a confederation which indicates that the agreements are less strict

CLARIN established official contacts with TERENA and had already various discussions with national IDFs (Se, Fi, Dk, Ge, NL, Hu), since it is very important to good understanding what the current trends are. Also discussions with eduGain and the Nordic federation will take place to understand the solutions they are choosing and their motivations. The harmonization has not been solved yet in a way that CLARIN can build on the solutions seamlessly and therefore CLARIN will need to define its requirements and interact with the respective initiatives. Whatever CLARIN will do, we need to take over the solutions that are already widely agreed upon at political level. Since almost all CLARIN centres are part of national identity federations already, we also need to know exactly what the various national circumstances are. We will describe the state and perspectives in a separate document.

2.4 Service Provider Federations

This term is fairly new, since until now service providers appear as individual companies acting according to a specific business model or within the IDFs to provide for example learning material. In our context, service provider federations are unions of research and other institutions that can offer services under the same set of coherent terms, i.e. they define license models, code of conducts, offer help facilities, carry out joint developments, share central services, specify the kind of user credentials they need for authorization purposes, etc. GEANT is a good example for such a federation although it is called differently, since the users are the research institutions, but they share the need in special network services, for which it makes sense to create separate institutions at national and at European level which are funded by the research world directly or indirectly.

In the centres document we already sketched a typical scenario that covers identity providers, national IDFs and two service provider federations. Every organization is free to join such a SPF as a member, since establishing and maintaining an SPF will cost money. Making use of the services of a SPF only makes sense for those institutions which want to give their researchers an opportunity to access the material, thus also being consumer of the services of an SPF needs to be a free decision.



This figure shows a possible scenario where two identity federations established a domain of clear rules about user management. One service provider (federation) obviously includes all members of the two IDFs as consumers. Another SPF just includes only some of the IDF members as consumers; however, they can rely on the same rules.

CLARIN is focusing on a federation that will integrate all relevant resource and technology providers primarily based on stable service centres. This note and the work in work package 7 need to work out in more detail what kind of agreements will be necessary to form such a federation of resource providers.

Also organizations outside of the research domain such as big companies, libraries and archives could make use of the services from the SPF and the SPF could make contracts to include, repackage and pass through services from the same group to its users. Here we are thinking of companies such as Microsoft and Google, the big national libraries and archives and many other which are not part of the academic world, but who have much data that is of use for the same group of researchers that CLARIN is addressing. Business models need to be added if access models are included that require some fees.

The benefits of establishing such SPFs can be summarized as:

- defining simplified and harmonized terms of licensing
- harmonizing the Code of Conducts associated with usage of the material or services
- in the ideal case a researcher does not need to sign any license agreements, since this was done between the SPF and the IDF the researcher is member of
- in the ideal case the researcher needs to sign a Code of Conduct only once, since the regulations are the same for all resource providers
- establishing a single sign-on/single identity mechanism for all language resources and technology allowing the researchers to create virtual collections crossing institutional boundaries and virtual applications where different services are combined to new powerful operations

A serious problem emerging for SPFs however has not yet been tackled and needs to be addressed by CLARIN: in general signatures under license agreements will be stored at the authorization (resource provider) side, however when a user agreed with the terms of a Service Provider Federation (SPF) the acceptance of a general license agreement or general code of conduct is not associated with a single resource provider, but it rather becomes an attribute of a user in his relation with the SPF. So, also this information needs to be accessible to prevent useless overhead for the users independent of which service provider of the federation was contacted first. Yet this problem has not been tackled so far by the federation or grid communities, but needs to be addressed within CLARIN.

2.5 Preservation Federations

Long-term preservation of digital data is amongst other issues very much dependent on a clear data distribution strategy, i.e. resources need to be copied and stored at various places. Since an archive in general will always ask the "right to archive" which includes the right to create copies at other locations for redundancy reasons, there is a legal basis for this. However, an archive will make agreements with depositors that specify the rules under which the deposited data may be accessed. Increasingly often these copies will not only be used as backup copies, but also for optimizing the access to them. Whatever the solutions are, they imply that the archive that will distribute the data needs to make formal agreements with the sites hosting the copies. Also data and information transfer protocols need to be specified etc. Thus we can speak of preservation federations built for specific tasks.

We can refer to two examples where such strategies are being applied. In the Chinese Digital Museum project university museums established a "Distributed, Standards-based Repository Federation" [Tansley 2006] which finally will include 100 universities. The task of this federation is to allow these museums to seamlessly replicate metadata and content from other members. All museums are using the DSpace repository system, for the data exchange standard protocols and container formats are used such as [PMH] and [METS] and a common domain for persistent and unique identifiers was established. This federation is based on a number of rules establishing a trust domain and a set of technology agreements.

In the [DOBES] project similar techniques are applied to preserve the material about languages and cultures that soon will become extinct for future generations. A core of 4 large computer centers is used to archive copies of all material and in addition an increasing amount of "regional repositories" has been set up in various countries to add a new dimension in preservation and active involvement. Currently 10 such regional repositories have not only copies of parts of the data, but also offer them as services. Also the copies at two of the large computer centers will be offered as services in near future. In all cases agreements determine which data can be copied to which other places and what kind of behavior rules are associated with this material. Also in this case we can speak about a federation which has as its primary function the long term preservation

of the resources, but as secondary function also to improve accessibility. In the CLARIN network of centres there will be a sub network of those centres offering data preservation services. They should be in that state to exchange data as described above for the purpose of long-term preservation.

2.6 Metadata Domains

In this context it makes also sense to refer to large domains where metadata is shared. Since metadata is open, one can only speak about shallow federations where it is more a point of good attitude to not misuse or modify the harvested information. Basically a few technological criteria are defined to participate in such domains. Frequently, compliance with the OAI PMH protocol and the additional delivery of records using the Dublin Core [DC] semantics are required. It is left to chapter 3.1 and another document to explain CLARIN's strategy in this respect.

2.7 Persistent Identifier Domains

Persistent identifiers for resources will become a key pillar in the emerging linked domain. There are different suggestions of how to achieve persistency of the references. The W3C [TAG] is advising to use http URIs where special caution with the construction of the URI should guarantee that it does not include changing components. Increasingly often domains for persistent and unique identifiers are created by communities that do not rely on the persistence of the chosen URIs. Even if we may belief that a well-chosen URI may be persistent, for others they just appear as any other URL with limited validity. This is the reason that increasingly often strong communities such as the group of national libraries [NBN], the Australian research and education domain [PILIN] or many of the publishing companies start establishing federations where they agree on a schema for PIDs and a registration and resolution mechanisms. Currently, we can identify at least two major suggestions in parallel to the usage of proper URIs: (1) URNs are suggested by the big libraries, but yet there is no convincing public registration and resolution system. (2) Handles are suggested by some initiatives and with the Handle System [HS] there is a widely used resolver. The latter system is also used by the big publishers who created a commercial service called [DOI] on top of the Handle System. In this case we can speak about a deep federation since for example the International DOI Federation was built to also make statements about the quality of service and a funding scheme. In other cases just an offer is made to the scientific community which can be accepted by individual members.

For CLARIN we need services that guarantee that an identifier associated with a resource will be maintained over a longer period of time. In the case of an intermediate layer as with URNs and Handles some agency needs to guarantee that an identifier can be resolved to the same resource over the same long period. Since there is no widely usable and non-commercial service yet CLARIN decided to establish a Handle System based service that can be used by all CLARIN members to register resources.

For further details about persistent identifiers we refer to the document CLARIN-2/2008.

2.8 GRID world

The term "Grid" used in relation to information technology was first applied in relation to combining high performance computers in a distributed manner to tackle the Grand Challenges, i.e. it was introduced with the denotation of "Grid Computing". So the computation itself is historically central to the Grids although the meaning of Grid has recently been extended to cover all the spectrum of things involved in computation, including data storage, transport and ownership, and also collaboration tools. The term "Grid" was in particular extended to "Data Grids" by this community to indicate that in distributed computing scenarios data needs to be transferred via high speed mechanisms to the locations where the computation will be carried out. In short, the Grid is about enabling distributed work across institutional and geographic boundaries, but computation being the focus of such infrastructure. The general vision in the Grid world is that sometime in the future people will be able to reap benefits of grid infrastructure the same way as it is now possible to plug electronic appliances into the power grid. The Open Grid Service Architecture [OGSA] contains a proper description of such an infrastructure.

Also in these distributed applications security naturally has to be taken serious. The OGSA specification document describes a number of points of concern: Authentication and authorization, multiple security infrastructures, perimeter security solutions, isolation, delegation, security policy exchange, intrusion detection, protection, and secure logging. Typical authentication software such as Kerberos in conjunction with Public Key credentials can be used to give users access to the Grid infrastructure. European projects

such as [EGEE], [PRACE] and [DEISA] and various national projects are devoted to work on Grid aspects. This community created various middleware products such as [GLite], [Unicore], [GTK][6] etc. with a number of components tackling the above issues. Yet with respect to the authentication of users they rely on user certificates according to the X.509 standard and a hierarchical LDAP system. DEISA is now starting to look at the possibility of integrating [SAML]-based exchange of credentials and then link up to the emerging national IDFs. This is work in progress. It should be noted here that CLARIN will not expect users to be certified, since this would not work for many years in the social sciences and humanities.

Recent incarnations of Grid middleware strive to support or incorporate web services in their architecture, yet we cannot refer to a seamlessly running implementation and in general CLARIN does not assume grid middleware being installed for accessing CLARIN's infrastructure. Therefore, a joint project between MPI/CLARIN and the Dutch Grid experts is working on a solution that will make it possible to execute web applications and web services in a federated authentication and authorization infrastructure (AAI) environment, i.e. making use of delegated authority when for example chains of operations are executed where the identity of the user needs to be preserved throughout the whole process. The aspect of allowing web services or in general applications to authenticate within a federation type AAI to accessing distributed resources has been discussed more deeply in the grid community and the Switch federation [SWITCH] created a technique called [SLCS] (Short Lived Credential Service). Currently, this technique is being implemented by [FEIDE] and [Surfnet].

It seems that currently the two communities, grid and digital libraries pushing the federation ideas, are coming closer together.

2.9 Implications for CLARIN

CLARIN needs to establish a Service Provider Federation for language resources and tools that can act like one big distributed publisher, i.e. it can sign agreements with Identity Federations and in doing so simplify the bureaucratic hurdles for the individual researchers. Since there will be access restrictions due to rights and in particular privacy issues and since there will be shared services at national and at European level (and even beyond), it will be necessary to specify agreements and rules that guide the CLARIN internal work and the access of the users to the offered services. It is the task of WP8 to define the political/organizational framework in which the technical solutions will be embedded and it is the task of WP7 to deal with all rights, license and ethical issues and with rules for access in detail. This document needs to describe the requirements for agreements and rules from the insights into the technical aspects of federations.

It is obvious that each academic institution will be part of several federations dependent on its scientific interests. To make such a multiple scheme feasible all SPFs and IDFs need to harmonize the way to define trust relationships. Therefore TERENA and EduGAIN[7] will play an important role and CLARIN will as much as possible link up with both. In particular it can be stated that CLARIN will adhere to the general trends in using attributes and values. Individual organizations or individual states may define their own schemas, but they need to be mapped to each other at the European level. Here the SCHAC schema seems to be a good point of departure.

Since central services are necessary, since certain services need to be associated with service quality statements and since rights and privacy issues are involved CLARIN will need to establish basically a shallow federation that does not include strong penalties, but it needs to have some elements of deep federations. Also the trust relation between the CLARIN SPF and the national IDFs needs to be based on a set of rules. It needs to be investigated which kind of soft penalties such as exclusion will be required. With respect to the scenarios presented earlier CLARIN will be faced even with the more complex ones. It will be one of the outstanding services to offer a search index on all metadata descriptions and resource content which is integrated into the CLARIN domain.

According to this scenario we can derive a few general rules which need to be looked at in more detail by WP7:

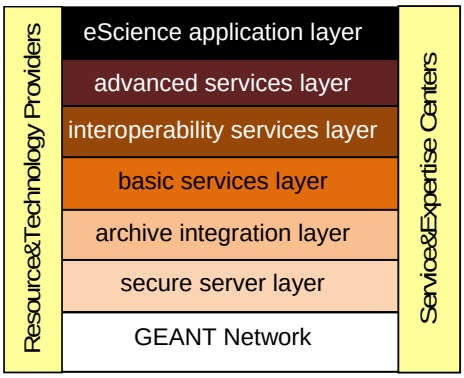
[6] It is generally known that all these products are not yet operating satisfyingly and stable enough to speak about a mature middleware software.

[7] TERENA and EduGain are working closely together so that the interaction can be organized efficiently.

- members of a Service Provider Federation share a mission and need a formal framework to take over responsibilities and to formulate a duty to adhere to a set of rules; they will establish domains of trust with Identity Federations who represent potential consumers of services
- user management needs to be audited by every IDF according to the trust agreements where CLARIN primarily has to accept the terms that are defined by the national IDFs and secondarily collaborate with TERENA/EduGain to make use of harmonized services and to specify and implement missing functionality
- in collaboration with TERENA/EduGain architectures need to be created that allow to filter and map schemas so that the required user attributes for granting access are being transferred
- access management is left to the institution that "owns" the resource or acts on behalf of the depositor where CLARIN will make great efforts in collaboration with others to simplify and harmonize license conditions and code of conducts
- the centers involved as service providers remain fully independent, but need to make statements about how long they will be able to give certain services and at which service quality level they will operate
- for all centers that will participate in the SPF the detailed schemas from the corresponding IDF need to be analyzed and harmonized

3. Federation Technologies

Establishing a federation of service centers requires setting up a distributed layer of middleware technologies. These are typically localized between the network and the application layers and do not include too discipline specific components. There is an overlap with some of the activities in the Grid community. In the following figure this layered system is indicated.



The figure gives a schematic view on a layered system of systems and responsibility. On top of the network provided by Geant mechanisms need to be implemented that allow a secure server interaction, a virtual integration of archives and that offer typical integration functionality such as handling persistent identifiers. These layers are dealing with "federation" aspects. On top of this we need to provide a layer that facilitates interoperability such as converters, concepts registries and ontologies. There will be a number of advanced services such as metadata and content search engines that make use of the interoperability layers. Finally there will be new types of applications that make use of all functionality offered by the other layers. These three layers are typically associated with the term "eScience".

In this document we summarize the following aspects under the heading "federation technologies":

- a system of servers and services that can interact in a secure way based on widely accepted and signed certificates
- a distributed system for authentication and authorization
- a distributed system of metadata providers, registration services and portals
- a distributed system offering services to register and resolving persistent identifiers for all types of resources including human resources
- a distributed registry for all types of resources such as participating centers

In the following we will discuss these services in more detail.

3.1 Secure Server Interaction

In many respects an infrastructure is depending on interacting servers and services as has been indicated. Given the utterly insecure state of the internet any interaction scheme that does not include verification of the servers identity would cause failure of all infrastructure plans. This problem is not new of course and it has been taken up by the grid community for example, so that CLARIN can rely on the chosen solutions.

[EUGridPMA] (European Policy Management Authority for Grid Authentication) is the European organization to coordinate the trust fabric for e-Science grid authentication in Europe. It is the European authority that is accepted to establish requirements and best practices for grid identity providers to enable a common trust domain applicable to authentication of end-entities in inter-organizational access to distributed resources. As its main activity EUGridPMA coordinates a Public Key Infrastructure (PKI) for use with Grid authentication middleware. EUGridPMA itself does not provide identity assertions, but instead asserts that – within the scope of this charter – the certificates issued by the Accredited Authorities meet or exceed the relevant guidelines.

To support this the TACAR (TERENA Academic CA[1] Repository) repository is maintained which is a trusted repository containing verified root-CA certificates that can be entered into local lists. The certificates to be collected are those directly managed by the member [NREN]s, or those belonging either to a National Academic PKI in the TERENA member countries (NPKIs), or those managed by institutions to support non-profit research projects that involve the academic community. Thus, for each European country (and beyond) there are authorities that can issue certificates based on a formal procedure. This document is not the place to describe the procedure in detail. The national authorities need to be approached. There will be courses of how to certify the servers. Given the formal procedure it is wise to contact national authorities as soon as possible.

For all CLARIN centers of Type A and B it is obligatory to obtain such certificates for their servers and to make the appropriate configuration settings at system level. In doing so in general all services provided by such a certified server can be validated as well.

3.2 Distributed Authentication and Authorization

In previous chapters we already introduced a number of different federation scenarios and types. In this chapter we want to briefly summarize the functional elements of an AA infrastructure, before describing in chapter 5 some technologies that are around to implement such an AA infrastructure.

Independent of the scenario and type we can formulate that the underlying purpose of a federation is to establish a domain of trust so that independent institutions accept assertions from each other about users that are known and that have authenticated successfully. For the user the benefit is obvious, since wherever he is, he will be able to access resources in the federation

- by using one single identity and
- by logging in only once (single sign-on which is not easy to achieve)
- by signing federation-wide service provider conditions only once

This does not mean that an authenticated user can access immediately all resources that could be available to him. Because it can be possible that such a user

- has to sign usage conditions first that are specific for the different centers
- is forbidden access to certain resources
- will have to pay fees for accessing certain resources

More important is the scenario where institutes trust each other with respect to the "role" a specific person has in his home institution. A person could be "director", "researcher", "student participating in a specific class" etc. Managing authorization could be very much simplified if the authorization record could just specify that "all researchers of university X" are allowed to access certain data. Increasingly more experts tend to argue that in an eScience scenario where researchers and students are getting used to accessing and combining various data resources from different providers this will be the only way to make access management feasible. The domain of trust should allow us to rely on assertions about the user's role.

In the following chapters we will refer to a number of essential aspects that need to be taken care of.

3.2.1 User Management

[1] CA = Certificate Authority; RA = Registration Authority

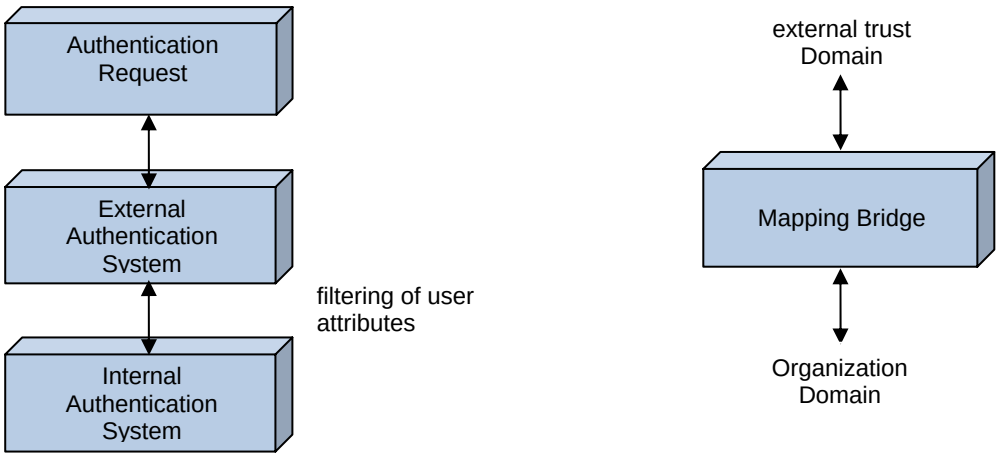
Most important in a domain of trust, is that institutes can rely on each other's user management. Rules need to be specified what kinds of attributes will be used and what kinds of values these attributes can have. The attributes typically address such issues such as "what is the status (role) of a person", "how long is the contractual relation", "which department does the person belong to" etc. The attributes and values (both together are often called schemas) need to be properly defined and need to be mapped to current practices in the different organizations, so that there is a common agreement what is accepted within the federation. It needs to be checked in how far audits are required to ensure proper user management.

There is a wide agreement to make use of the specifications from [EduPerson], [inetOrgPerson] and its implementation in RFC 2798[8]. However, culture and organization specific issues are not addressed. This is the reason why for example the existing European federations are using different schemas. It is the task of TERENA/EduGain to work out a harmonization and mapping strategy which already resulted in the SCHAC schema for example. Yet it needs to be shown in how far discipline-oriented domains such as CLARIN can make efficient use of the implemented solutions.

CLARIN as any other European infrastructure project can only rely on the activities of the national federations and the efforts of TERENA/EduGain. However, recent discussions with both have shown that community efforts such as CLARIN need to enforce priorities to come to smoothly operating solutions. Also, the fact that some CLARIN members do not have access to a national IDF, may force us to (temporarily) create a separate CLARIN user management policy.

Yet there are two other difficulties: (1) A research institute necessarily needs the freedom to handle user management in a flexible way so as not to hinder research by too much bureaucracy. (2) Large distributed institutions may have internal ways of communication where managing directors need access to confidential information that is forbidden to others. These two situations need to be resolved such, that at the one hand a domain of trust to the outside world is established and that also the needed flexibility in the internal domain is preserved.

To deal with the flexibility aspect each institute can implement a double layered scheme where the internal system for user management is used to access resources by all people known. A filtering process will create the second layer (see left figure) with all user entries that are compliant with the rules agreed in the trust domain. The internal system would cater for example for handling guests and research fellows in a simple way without hampering the trust relation to the outside world. At the filtering action one will include only those users that have a clear contractual status.



With respect to the second aspect similar arguments may hold. In an organization one may need a much more detailed system of roles which is not given by EduPerson for example. As depicted in the right figure a bridge could be used that does a schema mapping between the two domains. In Germany another aspect was encountered that may need to be looked at. From the outside world the Max Planck Society is treated as one legal organization. Due to its many different disciplines and institute cultures it makes very much sense,

[8] For RFC 2798 there is an existing LDAP schema that could be re-used.

however, to let the user management be done by the institutes. There are a few arguments supporting such a scheme: in general we are not allowed to send user passwords etc via the web to a central place, central user management would fail since they don't know the users, etc. This implies that for every authentication call the central organization instance will pass the request through to the local instance to carry out the authentication. The response of the local instance needs to be passed through again to the requester. Such a scheme is currently under development in Germany.

Each CLARIN centre needs to understand the framework in which it will operate in detail. At CLARIN level a Europe-wide exchange and correct interpretation of attributes is a prerequisite for a functioning infrastructure.

3.2.2 Attribute Exchange

As indicated we need to be sure that the interacting servers are who they claim to be. We already discussed the point of certifying the servers. Since the messages containing user attributes are crossing the internet via various ways we need to be sure that the messages cannot be intercepted by others by properly encoding them using these same certificates. Currently, there is a worldwide agreement to use SAML (Security Assertion Markup Language) as the message format.

SAML is an XML based language to exchange security-related information and provides functions to describe and transmit such information. SAML was developed by an [OASIS] consortium from 2001 including big companies. Supporting Single Sign-on was one of the applications in focus. Other applications were to support distributed transactions where several users collaborate on a transaction and share the security information and distributed authorization where another instance is authenticating the user. The current version SAML 2.0 is a kind of standard for identity federations.

SAML is defined in terms of assertions, protocols, bindings, and profiles. An *assertion* is a package of information that supplies one or more statements made by a SAML authority about authentication, attributes and authorization decisions. The *protocol* specifies the types of requests and answers that can be exchanged. The *bindings* specify how content can be exchanged using certain communication protocols such as HTTP/REST, SOAP etc. An important step forward was that SAML V2.0 permits attribute statements, name identifiers, or entire assertions to be encrypted. This feature ensures that end-to-end confidentiality of these elements may be supported as needed.

SAML 2.0 therefore is the basis for all CLARIN federation work.

3.2.3 Web applications/services

When talking about web access normally people have a scenario in mind where a user uses a web browser and tries to access a single resource per time which is indentified by a URL via that browser. The browser can store certain information in a cookie as long as the session remains active and it supports re-directs - a mechanism often used when authenticating in a distributed environment. CLARIN, however, is also interested in setting up an infrastructure where applications, web applications and web services on behalf of a specific user will request access to certain resources as well. Therefore all technical solutions need to consider all such scenarios where a service acts on behalf of a user (delegation) and can suitable credentials for that purpose. As already indicated the SWITCH federation created the SLCS solution which is currently being implemented in different countries with involvement of the MPI in a Dutch project.

In CLARIN we will need to find a solution to extend the distributed authentication and single sign-on principle to web applications and web services.

3.3 Metadata Framework

In chapter 2.6 we briefly introduced the importance of a joint metadata domain as a basis for an infrastructure. In a forthcoming document we will describe in detail how CLARIN can make use of the available experience to come to a new convincing solution that will allow us to realize a much broader coverage.

3.4 Persistent Identifiers

In chapter 2.7 we already stressed the need of introducing persistent identifiers. It can be introduced stepwise and CLARIN needs to offer a stable service for registration and resolving. Another document (CLARIN Persistent Identifiers, CLARIN-2/2008) from May 2008 describes the choices and requirements.

3.5 Centre Registry

As indicated in other CLARIN documents [CLARIN-1/2008, CLARIN-3-2008], dedicated centers will be the backbone of the CLARIN infrastructure. They will have various states and various functionalities within CLARIN. Similar to the Chinese museum project we will need a registry where we maintain a number of parameters for visual control and for machine interaction for all centers ensuring that the state of the infrastructure is under control. This registry needs to be stored in a redundant way to ensure availability and it needs to be protected against attacks.

This registry will have information for different functions that are not overlapping with the LRT registry. The following list may give an impression and needs further refinement:

- type of centre (see CLARIN Centers Types, CLARIN-1/2008, May 2008)
- certification information
- type of services
- base addresses for services
- state of services
- commitment information
- quality of service information
- etc

In particular the aspect of services needs to be studied with great care, since user oriented services, i.e. information about data resources and tools needs to be included in the LRT registries and since the description of the services will depend on the type of service. This registry needs to contain descriptions in particular about typical infrastructure services such as "this centre offers a PID service", "where are the mirror services", etc.

3.6 Experiences

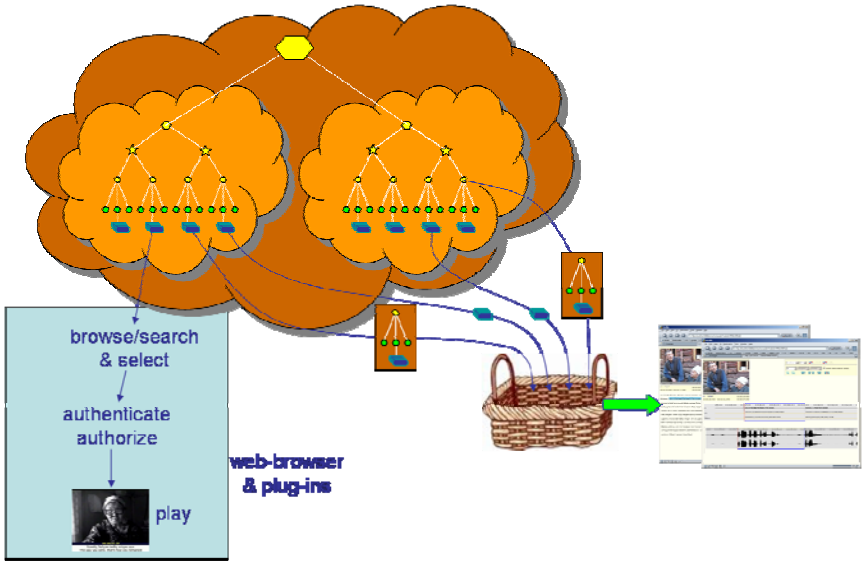
Here we want to describe some aspects that result from the experiences mainly gathered in the [DAM-LR] project, but also gathered from recent discussions.

- 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 project-based 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 integration. Changes, however, are not always easy since they may affect workflow processes etc.
- There is no off-the-shelf grid/federation 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 many of the typical components (Apache, Tomcat, IMDI, LDAP, Handle System, and 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 of the departments and institutions are not equipped with enough expertise to carry out the required installation and integration 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.
- At this moment we will probably lack a broad understanding at political level (university boards, institute directors, etc) about the general requirements put forward by establishing a research infrastructure.

- The investments to establish and maintain a federation are considerable and the investments in centers to run and maintain them as important infrastructure nodes cannot be neglected. It is obvious that university departments or institutions in general will not be able to maintain federation functionality with all its aspects over a longer period if there is no additional external expertise they can count on or if there is not a bundling of forces, for example, by a local collaboration between a computer centre and a department.
- We are lacking widely agreed standards at many aspects such as for example metadata schema, user credentials, federation agreement types etc. At the European level we can indicate that TERENA/EduGain are working on harmonization, however, yet this harmonization was not driven by communities. Priorities may have to be redefined to support a research infrastructure such as CLARIN.

4. AA Middleware

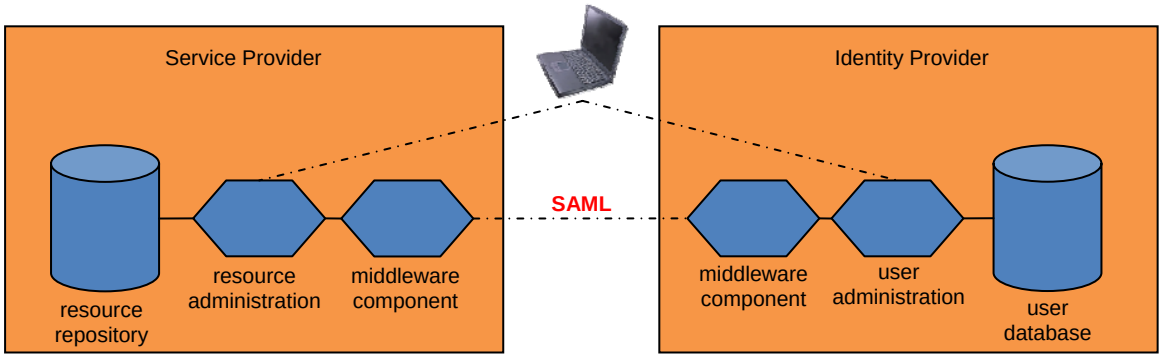
In the previous sections we explained the relevance of federations and indicated the type of technologies that are needed. The typical scenario that one wants to achieve is depicted in the following figure. Traditionally a user will use some search or browsing mechanism to find a single object, register him, obtain access permissions and access the object as indicated in the left part of the figure. In CLARIN the idea is that researchers can create for example a virtual collection by using (complex) resources from various archives with the goal to compare linguistic structure for example. To support the user in such scenarios we need to establish the single identity and single sign-on mechanisms.



The general architecture for such a scenario is specified by a few components:

- the user sitting at a notebook using a client application with help of which he wants to access resources
- a resource administration system that grants access permission to individuals or groups of users
- a user administration system that authenticates a user and
- a middleware that exchanges information between the two sides: the authorization and the authentication system
- A virtual collection (vc) registry mechanism, allowing the vc to obtain persistency and describe it accurately with metadata.

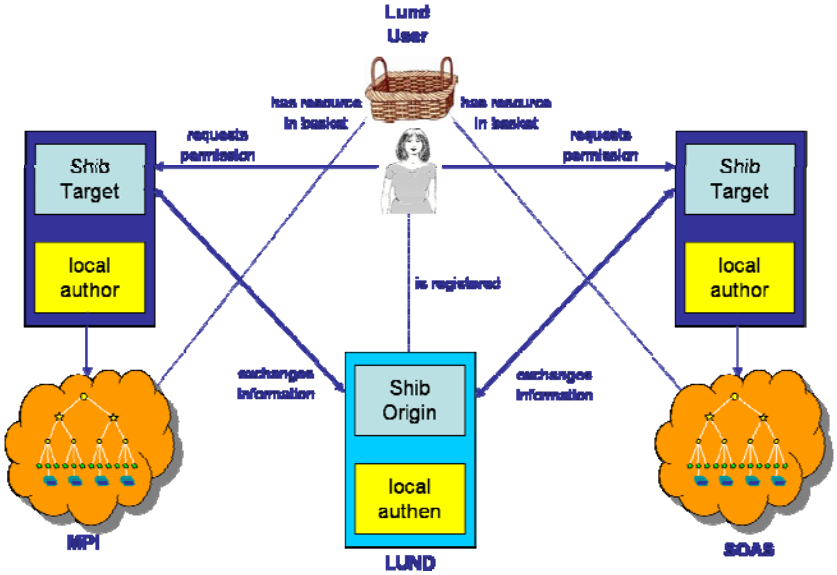
In general the middleware falls apart in two sub components: (1) one component is hosted at the authentication side (authentication provider) interacting with the local authentication system; (2) another component is hosted at the resource administration side (resource provider) interacting with the local resource administration. This scenario is shown in the figure above. It was already indicated that in CLARIN the interaction between the middleware components will be based on SAML and that we assume that the interacting servers are certified,



There are a few solutions that can be used in this scenario. Mostly [LDAP] and [ADS] are used to locally administer users, but also resource providers have developed own user management systems as part of the general information infrastructure.- it is obvious that these will not be directly supported in the shown scenario anymore. User management is expected to be carried out by trustful entities within an identity federation. The range of solutions for administering the access permissions to resources is large. Often Apache is used as a resource server, where the HT-access file or a DB equivalent is used to specify which user has which permissions to a resource, but often specific database systems with online frontends are used. With respect to the middleware we will restrict ourselves to introducing [Shibboleth] and [SimpleSAMLphp] as components that are widely used already.

Shibboleth

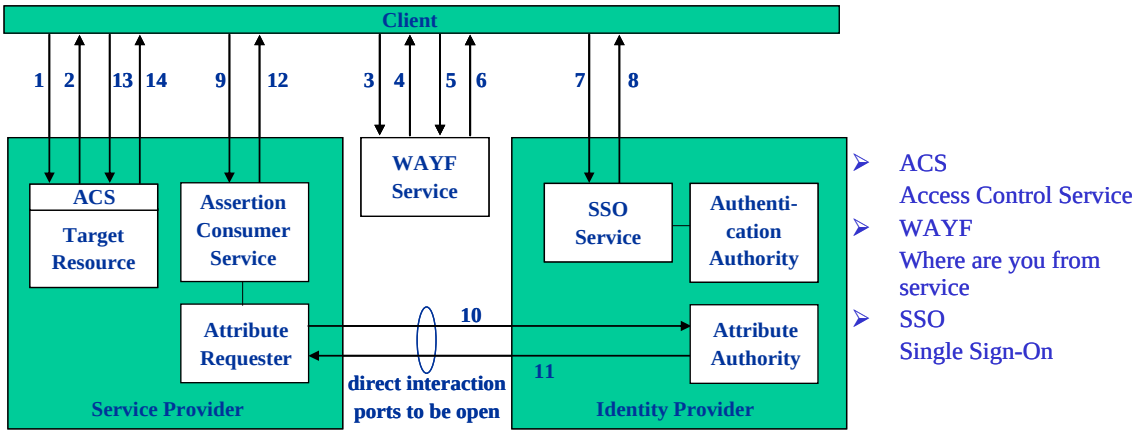
Shibboleth is a software product that was designed within the Internet 2 project to primarily facilitate distributed authentication in a scenario where groups need access and where group marks are exchanged. It was designed to help in the access scenario dominated by groups. It is increasingly often accepted by universities, libraries, publishing companies etc in various countries as a basis for a distributed authentication and authorization software, i.e. there is a broad user community. We can expect that institutions will increasingly often accept Shibboleth for the kind of trusted operations as required in distributed scenarios. There are two variants of Shibboleth, version 1.3 and version 2. Most still use Shibboleth 1.3 since it seems to be mature. Shibboleth 2 is supporting SAML 2.0 and comes with many more features; however, it still needs improvements.



This diagram indicates the principle function and interaction in a distributed authentication and authorization scenario. The drawing uses old terminology for the Shibboleth Service Provider (Target) and the Identity Provider (Origin) components.

The interaction is as indicated in the figure above. The user wants to access resources from a resource provider by using a URL in a web client. The Shibboleth Service Providing (SP) component knows that the resource is protected and requires an authentication. The authentication request is forwarded to the Shibboleth Identity Providing (IP) component at the home institution of the user. The home institution asks the user for username and password (or another mechanism) and in case of a successful authentication hands over the user attributes to the IP component. These attributes are passed through to the SP component and

then to the local access management software so that the user finally can access the resource. Since the user credentials can be stored by rewriting the original resource request URL, accessing the following resources could be carried out with less overhead. The following figure indicates that Shibboleth involves many HTTP redirects to hand over control to different entities. It uses an intermediate component call WAYF (Where Are You From)[9] to ask the user what his home institution is. At the end of the interaction process the two Shibboleth components establish a secure SAML based interaction to exchange the user credentials. Shibboleth[10] comes along with modules that can for example interact with LDAP at the identity providing side and with Apache at the resource providing side. A mapping needs to be defined, for example, to specify what LDAP attributes need to be handed over to the Shibboleth component. For other solutions than LDAP and Apache it is possible that modules need to be developed, i.e. any resource and identity provider needs to check in detail how his local solutions can interact with such middleware components.



1 Get Resource
2 Redirect (302)
3 Get Form
4 Send Form (200)
5 Submit Form
6 Send Cookie and redirect (302)
7 Request Authentication
8 Authentication Response
9 Send an Assertion Profile
10 Request Attributes
11 Send Attributes
12 Redirect with attributes
13 Send attributes for check
14 Provide Resource

Within several projects such as in DAM-LR this solution was implemented. It became obvious that still Shibboleth is not easy to integrate. It requires thorough knowledge and experience of experts to carry out the integration with the local AA components. The parameter files that are required need to be adapted to the user attributes used at both sides for example. A successful integration requires a clear strategy for the local AA components and of course agreements about the attributes and values used - both being actually independent of the middleware used.

SimpleSAMLphp

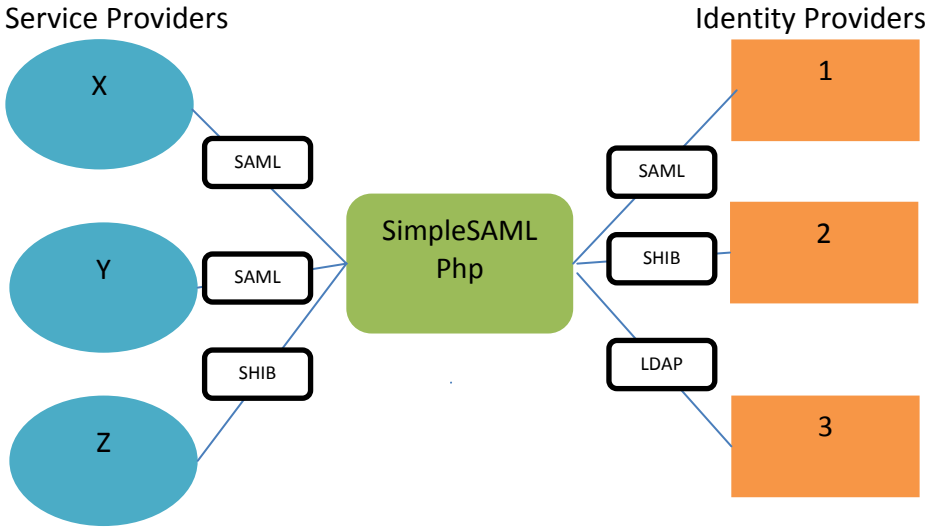
Another middleware component, SimpleSAMLphp, that can play a role in a distributed AA scenario was developed within the Norwegian identity project FEIDE and is now being used by several Nordic federations. It is an application that enables the low-barrier setup of an Identity Provider or Service Provider within a federation. Next to that, it can interconnect several Identity Providers and Service Providers using SAML 2.0 as a common standard. In practice this provides compatibility in heterogeneous AAI environments and even the bridging of multiple Federations (as in the case of EduGAIN).

[9] Increasingly often federations try to bypass the additional WAYF selections step by offering user realms, the domain the user is originating from is immediately provided (as with emails).
[10] Shibboleth 2.0 comes along with many more adapters to various known user management solutions such as for example relational databases.

The basic idea behind SimpleSAMLphp is to keep things more simple than with Shibboleth and to support a broad range of IP and SP solutions that are all based on SAML2.0. By the use of plugins, SimpleSAMLphp can interact with the following components:

- Shibboleth 1.3 and 2.0
- OpenID
- [A-Select] (Dutch IDP system), [CAS] (Danish IDP), [PAPI] (Spanish IDP)
- WS-Federation (several commercial providers: IBM, MS, Novell,…)
- Authentication via [PostgreSQL], [Radius], PKI (in progress)

SimpleSAMLphp therefore is an option to interlink different solutions that have been chosen already by different identity federations. It is said to be much easier to be installed and configured. The following figure indicates a typical scenario where SimpleSAMLphp plays an intermediating role connecting to various SAML talking components. As can be seen it also has an interface to LDAP if it is used to local user management. Typically SimpleSAMLphp is therefore used as intermediating component or as a component at the identity providing side.



CLARIN will need to be open to any middleware component that is introduced by one of the partners as long as it supports SAML2.0. However, it is the task of any centre to carry out the adaptation to its local environment.

5. Requirements

In this chapter we are summarizing the requirements for the emerging CLARIN infrastructure as they emerge from federation technology[11].

- CLARIN will establish a shallow federation in so far that no penalties will be specified except exclusion from participation in case of an inappropriate attitude. Every centre except for the infrastructure servicing centers can stop participation, however, it is expected that the services will be handed over to another centre by maintaining the stable references.
- Infrastructure centers need to make statements about the duration of their infrastructure service and the quality of service which can be expected. In this case there is the expectation of high availability of the service.
- Resource and service provider centers also need to indicate the duration and quality of their service.
- Every centre will certify its servers and services according to the accepted TERENA TACAR list. Special cases need to be evaluated.
- Every partner will setup a PKI system and sign its certificates with public keys.

[11] Requirements with respect to metadata infrastructure are discussed in other documents.

- A Handle System service will be offered to all centers to register, manage and resolve unique and persistent resource identifiers; other schemes can be used for referencing of course, but CLARIN will not offer services. A few centers will be required to set up mirror services. Every centre can of course manage its own Handle server which gives more freedom in assigning postfixes.
- CLARIN may develop services on the basis of the HS (like DOI), it is clear that organizations choosing other PID schemes cannot interoperate with them. CLARIN will not invest in building PID harmonization solutions.
- The authorization information for resources is exclusively maintained by the originating institution - this right is not touched by the AAI.
- It is left to the centers which kind of local solutions for user management they chose. However, they need to ensure that a proper interaction with either Shibboleth or SimpleSAMLphp takes place. No special mechanisms beyond username/password authentication are required by the CLARIN Service Provider Federation. It is left to the national IDF which level of user authentication they expect.
- It is left to the centers which kind of local solutions for resource management they chose. However, they need to ensure that a proper interaction with Shibboleth or SimpleSAMLphp takes place.
- As protocol for all AA interaction SAMl2.0 is chosen for CLARIN.
- A CLARIN WAYF needs to be setup and maintained.
- The user attributes are determined by the national identity federations. The CLARIN infrastructure will use the SCHAC scheme as a scheme for interoperability where necessary and integrate with the eduGAIN efforts where possible.
- The quality of user management is handled by the national identity federations at first instance. TERENA should establish a Europe-wide monitoring system for problematic cases. Auditing will not be required by the CLARIN Service Provider Federation. However, if CLARIN acts as a reseller or publisher of library services, they may need to require auditing. Also national IDPs may choose to audit.
- An extension of the single sign-one principle to web services needs to be worked out.
- An investigation of how to maintain SSO in the face of resource provider storage of licenses and "usage conditions" forms should be done.
- The elements of a registry of centers need to be worked out in detail.
- Hands-on workshops for Shibboleth and SimpleSAMLphp need to be organized.
- Help needs to be organized to help centres to setup the federation pillars.

6. Procedure

The procedure needs to be closely related with the selection of the centres. All setups are of temporary nature in the preparatory phase, i.e. we will not claim that the solutions will be continued in exactly the same way in the construction phase. In the preparatory phase we will not take care of redundant services. Currently we have about 25 declarations of interest to become a CLARIN centre[12]. We expect to have two rounds of establishing a network of centres: one in 2009 and an extension in 2010.

- Preparation & Help (this work already started and will be intensified at the beginning of 2009)
    - get selected centres to create a proper repository system
    - get selected centres to define their services dependent on the type
    - get selected centres to ask for certificates and setup a PKI system
    - motivate centres to start issuing PIDs according to the Handle System
    - organize help for federation setup
    - check the state of the harmonization efforts and solutions from eduGAIN
    - specify the way CLARIN-wide signatures should be handled
- Training (courses will be offered in the first half of 2009)
    - organize a training courses on federation technology (SAML, Shibboleth, SimpleSAMLphp)
    - organize a training course on proper user management and solutions (ADS, LDAP, Filtering)
    - organize a training courses on PID technology (Handle System, integration in Metadata)
    - organize a training course in PKI and server certification
- Implementation (this work already started and will be continued in 2009)
    - setup a central Handle System offer

[12] This is changing since new statements of interest will come in during the whole preparatory phase.

- developing some basic PID services such as adding checksums[13]
- get selected centers to install Shibboleth, SimpleSAMLphp or another SAML 2.0 compliant solution for IDP and SP
- setup a separate IDP for CLARIN to integrate those who don't yet have a national IDF
- specify the registry of centres and set it up
- create a WAYF registry (perhaps re-use the centres registry)
- ask centres to set up their IDP and SP components and to integrate them with their environment dependent on the situation (national IDF, local authentication and authorization solutions)
- make use of TERENA/eduGAIN harmonization efforts and adapt where necessary
- develop an AAI solution for web-services

- Non-Technical (this work will start in January 2009)
  - establishing technical and non-technical SPF requirements and agreements in relation with WP7

7. References

Projects and abbreviations

[ADS]	Active Directory Service	http://en.wikipedia.org/wiki/Active_Directory_Service
[A-SELECT]		http://a-select.surfnet.nl/
[BOAI]	Budapest Open Access Initiative	http://www.soros.org/openaccess/
[CAS]	Central Authentication Service	http://en.wikipedia.org/wiki/Central_Authentication_Service
[DAM-LR]	Distributed Access Management for Language Resources	http://www.dam-lr.eu/
[DC]	Dublin Core	http://dublincore.org/
[DEISA]	Distributed European Infrastructure for Supercomputing Applications	http://www.deisa.eu/
[DOBES]	Dokumentation Bedrohter Sprachen	http://www.mpi.nl/dobes
[DOI]	Digital Object Identifier	http://www.doi.org/
[EduGAIN]	GÉANT Authentication and Authorisation Infrastructure	http://www.edugain.org/
[EduPerson]		http://middleware.internet2.edu/eduperson/
[EduRoam]	EDUcation ROAMing	http://www.eduroam.org/
[EGEE]	Enabling Grids for E-sciencE	http://www.eu-egee.org/
[e-IRG]	e-Infrastructure Reflection Group	http://www.e-irg.eu/
[EUGridPMA]	European Policy Management Authority for Grid Authentication in e-Science	http://www.eugridpma.org/
[Federation]		http://en.wikipedia.org/wiki/Federation
[FEDORA]	Flexible Extensible Digital Object Repository Architecture	http://www.fedora-commons.org/
[FEIDE]	Identity management system on a national level for the educational sector in Norway	http://feide.no/
[GEANT2]		http://www.geant2.net/
[GLITE]		http://glite.web.cern.ch/glite/
[GTK]	Globus ToolKit	http://www.globus.org/toolkit/
[HS]	Handle System	http://www.handle.net/
[inetOrgPerson]		http://tools.ietf.org/html/rfc2798
[KALMAR]		http://rnd.feide.no/content/kalmar-union
[LDAP]	Lightweight Directory Access Protocol	http://en.wikipedia.org/wiki/Ldap
[METS]	Metadata Encoding and Transmission Standard	http://en.wikipedia.org/wiki/METS

[13] The effort possible here is dependent on the available person power and cannot be answered yet.

[NBN]	National Bibliography Number	http://en.wikipedia.org/wiki/National_Bibliography_Number
[NREN]	National research and education network	http://en.wikipedia.org/wiki/NREN
[OASIS]	Organization for the Advancement of Structured Information Standards	http://www.oasis-open.org/
[OGSA]	Open Grid Services Architecture	http://www.globus.org/ogsa/
[PAPI]	Point of Access to Providers of Information	http://papi.rediris.es/
[PILIN]	Persistent Identifier Linking Infrastructure	https://www.pilin.net.au/
[PMH]	Protocol for Metadata Harvesting	http://www.openarchives.org/OAI/openarchivesprotocol.html
[PostgreSQL]		http://www.postgresql.org/
[PRACE]	Partnership for Advanced Computing in Europe	http://www.prace-project.eu/
[RADIUS]	Remote Authentication Dial In User Service	http://en.wikipedia.org/wiki/RADIUS
[SAML]	Security Assertion Markup Language	http://en.wikipedia.org/wiki/SAML
[SCHAC]	SChema Harmonisation Committee	http://www.terena.org/activities/tf-emc2/schac.html
[Sibboleth]	Shibboleth	http://shibboleth.internet2.edu/
[SimpleSAML]	SimpleSAMLphp	http://rnd.feide.no/simplesamlphp
[SLCS]	Short Lived Credential Service	http://www.switch.ch/grid/slcs/
[SRU]	Search/Retrieve via URL	http://www.loc.gov/standards/sru/
[Surfnet]	Dutch NREN	http://www.surfnet.nl
[SWITCH]	Swiss NREN	http://www.switch.ch/
[TACAR]	TERENA Academic CA Repository	http://www.tacar.org/
[TAG]	Technical Architecture Group	http://www.w3.org/2001/tag/
[TERENA]	Trans-European Research and Education Networking Association	http://www.terena.org/
[UNICORE]	UNiform Interface to COmputing REsources	http://en.wikipedia.org/wiki/UNICORE

Literature

[Atkins 2003] 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, http://www.communitytechnology.org/nsf_ci_report/

[David 2006] David, P. A., den Besten, M., and Schroeder, R. (2006). How open is e-science? In *Proceedings of the IEEE 2 ndInternational Conference on eScience and Grid Computing*, pages 4-6.

[Leenars 2005] M. Leenars (2005). e_infrastructures Roadmap: http://www.-e-irg.org/roadmap/eIRG-roadmap.pdf

[Tansley 2006] Tansley, R. (2006). Building a Distributed, Standards-based Repository Federation. D-Lib Magazine , 12 (7/8), 1082-9873, http://dx.doi.org/10.1045/july2006-tansley

[Taylor 2001] 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

[Volanis 2006] Volanis, N. and Dumortier, J. (2006). A European Legal Approach to Grid Computing. In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*. IEEE Computer Society Washington, DC, USA.