GES 2007 German e-Science

Available online at http://www.ges2007.de

A Federation of Language Archives Enabling Future eHumanities Scenarios

P. Wittenburg, D. Broeder, M. Kemps-Snijders, A. Dimitriadis, Th. Soddemann

MPI für Psycholinguistik,
Wundtlaan 1, 6525 XD, Nijmegen, Niederlande
Email: {peter.wittenburg,daan.broeder,marc.kemps-snijders}@mpi.nl
alexis.dimitriadis@let.uu.nl, soddemann@rzg.mpg.de
phone: +31-24-3521113

Abstract

This paper describes the need for new infrastructures for future eScience scenarios in the humanities. Three projects working on different aspects of these infrastructures are examined in detail. The first project is trying to achieve a federation of archives, developing an integration layer at the level of localization, access to and referring to an archive's raw data objects. The other two try to achieve interoperability at the level of semantic interpretation of linguistic data-types and tagging systems. The project's different approaches to this problem show the trade-of between flexibility and the user's workload. All three approaches give an impression about the necessary steps to come to an eHumanities scenario.

1 Introduction

According to John Taylor[1] eScience is “about global collaboration in key areas of science, and the next generation of infrastructure that will enable it”. It means a new form of aggregation (1) in the way collaboration is carried out, since science is based on an open exchange of competing ideas and intensive scholarly interaction, (2) in such a way that the current existing boundaries for accessing a common domain of resources are overcome and (3) enabling new opportunities for cross-discipline fertilization. The realization of this eScience vision for the humanities, and in particular in the linguistic discipline will only succeed if scholars and students have seamless and sustained access to large distributed, but nevertheless virtually integrated repositories of useful language data, processing and knowledge resources. This holds both for “large scale problems”, such as giving more comprehensive answers to the question of how the human mind processes natural language, where many of the available linguistic resources have to be used to get new insights, as well as for the typical “small scale problems” where a researcher is looking for information for a comparatively modest inquiry, nevertheless using electronic resources housed at different locations.

[1] UK eScience: http://www.rcuk.ac.uk/escience/documents/report_coreproggrid.pdf

2 A Federation of Language Archives Enabling Future eHumanities Scenarios

The current infrastructure alone, which consists mainly of high-speed data networks and Internet services such as email and the World-Wide-Web, will not be sufficient to let all the expectations associated with the term “eScience” become reality. New type of enabling infrastructures are needed, such as those indicated by the term “Grid” which was extended to "Data Grids" addressing the matter of virtual data integration.

In the natural sciences, data exchange and integration problems are mainly concerned with the sheer mass of data and its data encoding formats. In the humanities, the major obstacle to data interoperability is format and semantic heterogeneity. Roughly speaking, it is in particular the differences in terminology that make it so difficult to cross the boundaries in our field and create a joint domain of language resources that can be utilized seamlessly.

In this paper we will address two infrastructures currently being worked out in concrete projects: (1) the first is about a Grid type integration of the collections of a few language archives (federation of archives) in the EU-funded DAM-LR (Distributed Access Management for Language Resources) project; (2) the second is about advanced web applications in the area of typology and corpus linguistics that make use of ontological knowledge to overcome semantic heterogeneity and that can be integrated in a flexible service oriented architecture. The solutions found and the suggested integration are building blocks for an eHumanities infrastructure.

2 Federation of Archives

DAM-LR is a EU project of four partners:[2] that have each an archive housing (digital) language resources. The aim of the project is to establish a federation[3] of archives, offering users a single virtual domain of resources, by adressing four main requirements with respect to interoperability and services:

- an integrated metadata domain that allows users to browse and search in a federation-wide metadata catalogue and to create their own work space by selecting resources from the various archives of the federation;
- a single domain of resource references where each resource is identified by a persistent unique resource identifier mappable to URLs via a robuste and highly-available resolution service. This

[2] Centre for Language and Literature. Lund University, Institute For Dutch Lexicography, Max-Planck-Institute for Psycholinguistics Nijmegen, NL, School of Oriental and African Studies, University of London

[3] In this paper we will not discuss the organizational, juridical and ethical dimensions that are also of great relevance when forming a federation of archives.

should also allow for efficient access to a resource when multiple instances are held across different federation sites;

- a single user identity accepted by all federation members and supported by a single sign-on system so that users only need to authenticate themselves once when they access resources at different members' sites during the resource selection process;
- an authorization system is needed that allows archive managers to give federation-wide access to users and groups to all copies of the resource within the federation.

This is all based on a domain of trusted servers and services – each service has to be able to prove that it is the one that it claims to be. As a means of implementing such a trusted domain, the TACAR[4] list of mutually agreed certificates was created, based on the principles of EUGridPMA[5]. In this implementation, national bodies declare that they will accept certificates from each other, with a Public Key Infrastructure used to sign certificates.

With respect to metadata interoperability the IMDI[6] metadata infrastructure is supported for browsing and searching either by using stored IMDI metadata or by creating them on the fly from a local format. IMDI was chosen because it supports not only resource discovery via searching and browsing, but also can be used for resource management which is regarded to be an essential function within federations. Several portals will be made available with full functionality for metadata search and browsing. For harvesting two methods can be applied: the OAI PMH[7] protocol, or direct harvesting of IMDI XML metadata via the browsable structure.

The interoperability requirement is the creation of a unified domain of unique resource identifiers (URIDs) to provide a stable method for referencing electronic resources. There are many reasons for introducing URIDs such as persistency over time, independence of the resource location, uniquely identifying both a resource and all its other instances and the possibility to resolve the identifier to all its multiple copies being stored at different locations. Conceptually URIDs can be compared with ISBN numbers that are used to uniquely identify published books. The federation partners chose the widely used Handle System (HS) to create, maintain and resolve URIDs. The HS fullfills all the aboved name desiderata. A URID or handle consists of a centrally issued prefix that uniquely identifies a local domain of authority that can freely issue posfixes, that form the second part of the handle thus delegating responsibility and guaranteeing the handle's uniqueness.

[4] TERENA Academic CA Repository

[5] European Policy Management Authority for Grid Authentication

[6] ISLE Metadata Initiative

[7] Open Archives Initiative, Protocol for Metadata Harvesting

The federation has agreed that all sites should respect and follow the access rights specified in the resource's authorization records as defined by its originating archive. This requires the authorization records to be available to all and since (in the DAM-LR setup) the availability of a resource depends also on the HS resolving the URID in the first place, placing the authorization information with the URID in the HS, something the HS supports, looks like a good choice. The Handle System records will be redundantly stored at multiple sites, but only the originating member will have full control of the handle records.

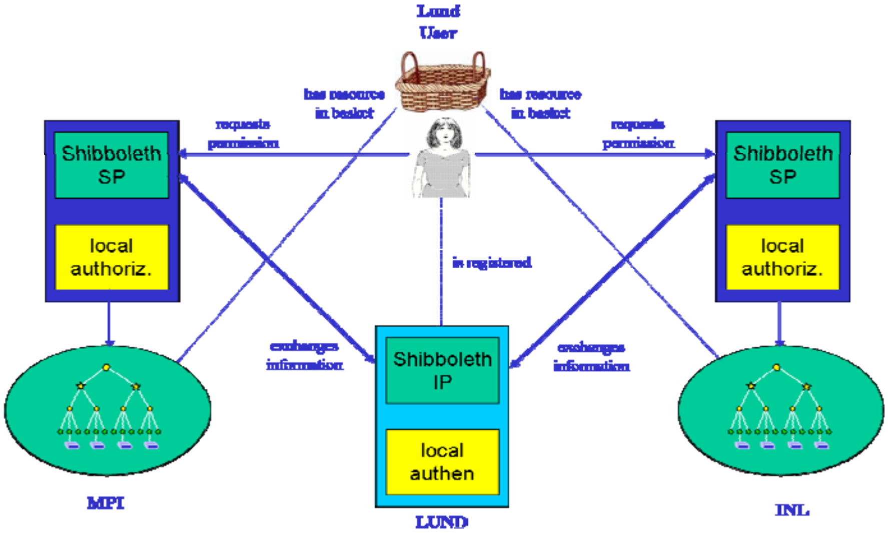


Figure1: Here a typical access scenario is shown where a user from Lund University may want to access data two other institutes and where Shibboleth plays the role of exchanging user credentials via secure mechanisms after the user was authenticated at his home institution.

With respect to authentication, federated user identity management and authorization the situation is more complex. One widely used contender for implementing these, Shibboleth, is excellent in circumstances where authorization policies can be described by statements such as "resource A is accessible by all members of university class students" and the number of groups or attributes shared by the federation members is small and its their semantics are undisputed. The authentication of the student is left up to the home institute and the resource originating institute grants access based on the attributes, verifyable with the students home institute, that specify his class membership. For researchers operating autonomously, as will often be the case for users in our domain, using Shibboleth requires an extra federation wide unique user ID attribute since granting access on the basis of group membership only is not possible. This also complicates the maintenance of the authorization records since these now contain identifiers of individual users and need to be regularly audited since user identifiers are more volatile then group attributes. The fact that Shibboleth has already received wide acceptance in the Digital Library domain influenced our choice, despite this inefficiency.

The federation partners agreed that user management will be performed at the home site and that only limited information about users will be exchanged. The prototypical system will support Open LDAP for the archive's local user management since it has many useful features; it is already widely used in the academic world and can be easily connected to Shibboleth. In addition, However each federation partner is free to setup their own authentication system if a connection to the Shibboleth identity provider can be provided..

A few components need to be added to complete the architecture. Firstly, we need an Access Control System that guards protected resources (that are in principle all accessible through ordinary HTTP requests) and forces authentication and authorization via Shibboleth. The Shibboleth Identity Provider component, that will take care of authentication of users, will be implemented as a TOMCAT container using a JAAS[8] realm. In this way a fallback mechanism can be created for authentication that tries the authentication mechanisms of different organizitional units. The actual protection of resources is done by the "standard" Shibboleth apache web-server add-on module "mod-shib". Finally, a management system that allows archive managers to efficiently manage the archive user records and set policies and permissions for their access to the resources will be added.

Looking at the future we can say that the integration layer that is being realized in the DAM-LR project with a few additional extensions can be used in scenarios that are dominated by web based services and web applications as it is foreseen in a next phase.

3 Crossing Semantic Boundaries

Crossing institutional boundaries as in the described archive integration project is not sufficient for creating an integrated domain of linguistic resources with different linguistic resource types such as annotated media, lexica, sketch grammars etc created in a wide variety of projects. Designing and developing integration frameworks to overcome the barriers created by different formats and structures and in particular by the usage of different terminologies has already a history in the humanities. We can refer to projects working on tagging and metadata standards, such as TEI[9], IMDI[10] and Dublin Core[11], to design generic models such as LMF (Lexicon Markup

[8] Java Authentication and Authorization Service

[9] Actually, TEI has existed for quite a while, but only recently received the attention that is necessary to achieve a higher degree of unification at the encoding level. http://www.tei-c.org/

10 IMDI is a metadata standard derived within the European ISLE project, http://www.mpi.nl/imdi

[11] Dublin Core: http://dublincore.org/

Framework)[12] and to work out recommendations for linguistic encoding with the help of Data Category Registries (DCR), such as within ISO TC37/SC4[13]. With the exception of metadata interoperability all suggestions have yet failed or are so new that it will take some time until tools will support them and that linguists will get used to them.

Although structural interoperability is still difficult to achieve we will focus in this paper on two projects started in the Netherlands that tackle the semantic interoperability problems with bottom-up, data driven approaches. The intention is to enable unified access to a variety of structured or semi-structured legacy resources such as typological databases, lexica and annotated media resources. At the MPI for Psycholinguistics a language resource archive is being maintained that contains more than 250.000 objects, mostly annotated media resources, but also lexica etc. These resources have been created by many different researchers who worked in many different projects. Due to ongoing projects this archive is continuously being extended in various ways, partly by enriching or correcting the existing linguistic annotation, partly by adding completely new collections. All researchers and projects are independent in their choice of how linguistic phenomena are encoded. Therefore, we cannot speak of a restricted domain of semantic concepts but of an open unbounded domain where often new concepts are introduced, and also where names that are already used in the archive can be reused to express different meanings, and where different terms are used to identify the same or rather similar meanings. The question that was tackled is which semantic interoperability mechanisms have to be available for researchers when they want to carry out, for example, searches across several of these collections at the same time.

The goal of the TDS (Typology Data Service) project[14], is to provide integrated access, through a web-based service, to a virtually integrated domain of typological databases, each containing a very large number of data fields (typically several hundred) about a large number of languages (again in the hundreds). Also here we can state that the typological databases were created independent of each other, with a focus on different aspects of languages and with different intentions in mind. Similar issues of semantic

[12] LMF: http://estime.spim.jussieu.fr/~pz/lrec2006/Francopoulo.pdf

[13] ISO TC37/SC4: http://www.tc37sc4.org/

[14] TDS is a project of the Netherlands Graduate School of Linguistics (LOT) including University of Amsterdam, Leiden University, Radboud University Nijmegen, and Utrecht University.

differences and similarities arise in the MPI case. However, in the TDS case we are dealing with a relatively small number of resources (databases), each semantically complex. There are about a dozen databases in the current initial phase, and the eventual size of the archive will be in the dozens rather than thousands of databases. Hence the focus was on unifying the semantics and encoding of a particular (but progressively extended) set of databases, i.e., the semantic scope of all concepts that are used within an initial set of typological databases could be carefully studied and analyzed. Again the question was which semantic interoperability mechanisms should be made available to the researchers to provide cross-database operations.

3.1 MPI Ontology Framework

In the MPI case a flexible ontology editing framework was designed to be the focus of the developments. It allows users to (1) easily extract the tags and values that are used in an actual selection of resources by using machine readable concept profiles (official concept definitions provided with a collection) where available; (2) easily select and combine linguistic concepts in personal concept registries; (3) easily link such personal concepts with concepts in central registries such as from the ISO Data Category Registry; (4) easily create personal relation registries that contain typed relations between concepts to be found in personal concept registries. Also concept definitions can be easily linked to the corpora where they are used and can be inspected in situ.

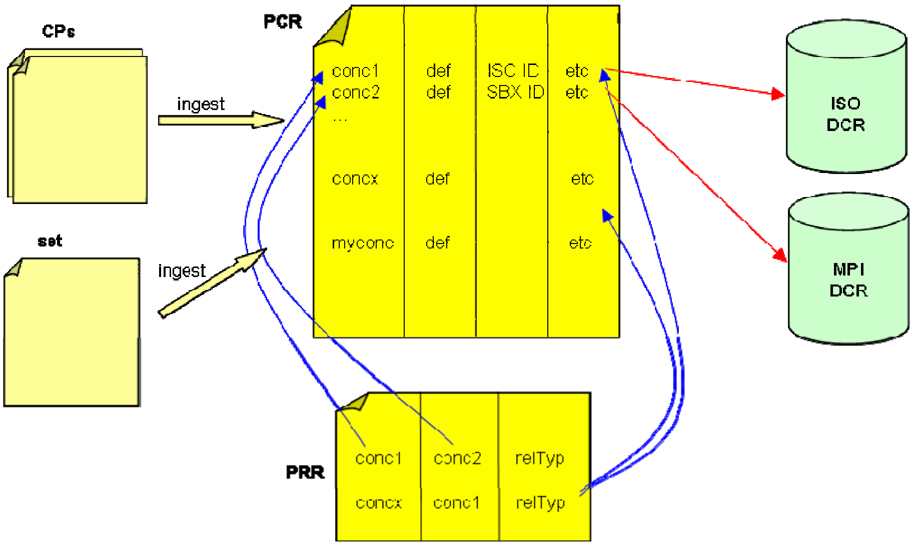


Figure 2: The most relevant linguistic data types to allow users to achieve semantic interoperability. In particular, users can create, manipulate and share personal registries for concepts and relations that can be used by search engines.

In figure 2 some of the linguistic data types are shown that are relevant. Concept Profiles (CP) can be associated with collections that are stored in the archive. The collection depositors are responsible to formally describe the concepts they are using and when selecting a resource an archive crawler will find the corresponding CP to allow the resesarcher to include its definitions. For most legacy resources the set of used concepts needs to be created on the fly by scanning the document, so that only a list of names can be presented to the user. Both can be included in Personal Concept Registries (PCR) that can be stored, manipulated and shared with others. The user may want to add references to central reference Data Category Registries. If two entries refer to the same DCR entry semantic equivalence can be assumed which can then be exploited by search engines for example. In Personal Relation Registry (PRR) the user can specify relations that are useful for his research activities and that also can be used by search engines. Also PRRs can be stored, manipulated and shared.

Only such a framework that is extendible and easily adaptable to the actual needs of the linguists will meet the needs. It can be expected that over time the amount of links to central concept registries will increase offering interoperability for free and that an increased number of useful knowledge modules will be openly registered making it easy for linguists to adapt them to their needs. To facilitate this work a first version of an editor was developed that allows users to manipulate the personal registries.

3.2 TDS Ontology Framework

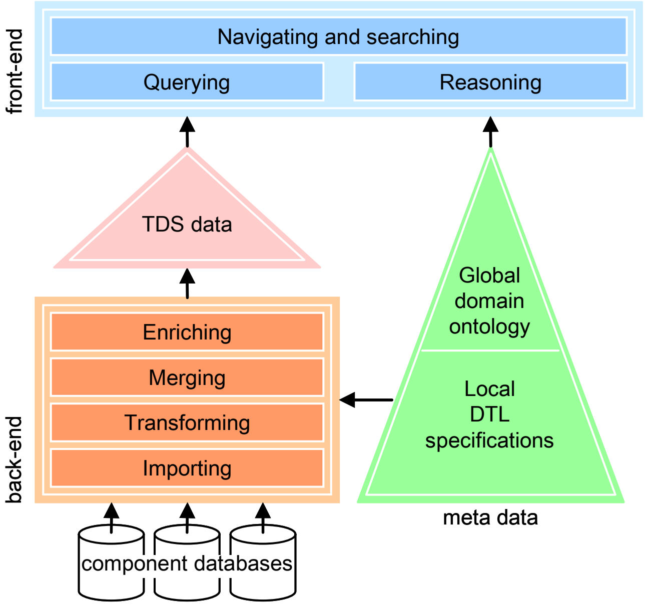


Figure 3: Component databases are integrated into the domain by complex transformation, merging and enriching processes. A two level ontology approach is provided to create a global domain ontology covering all included typology databases. While querying this domain ontology is used by the search engine to find corresponding entries in the included databases.

In the TDS case a complementary approach was chosen (cf. Figure 3). Based on an analysis of the databases to be integrated the TDS Ontology (TDSO) was designed so that it can be dynamically extended. It provides a non-prescriptive, "inclusive" framework of linguistic concepts and terms, into which the particular perspective of each component database can be integrated. Explicit links between the unified data and ontology concepts facilitate searching through the integrated databases and interpreting the found data. Searching is a two-step process. First, the user discovers fields relevant to the topic being researched, by using one of the searching or browsing options provided by the TDS interface. Selected fields are accumulated, forming a pre-query. In the second step, the user refines this pre-query and executes it.

The system relies on a *hybrid,* or *two-level,* ontology design: At the top is the global TDS Ontology of Linguistic Concepts which is a hierarchical data structure making use of OWL. It is extended by the database-specific *local ontologies*, which include the idiosyncratic definitions applicable to each database; the local ontologies are an integral part of the mappings between the databases and the TDSO, and are defined in the special-purpose Data Transformation Language (DTL) developed by the project. A DTL specification describes notions and their relationships, and the nodes in the tree are thus instantiations of these notions. To facilitate the work of the user different views of the TDS Ontology are offered.

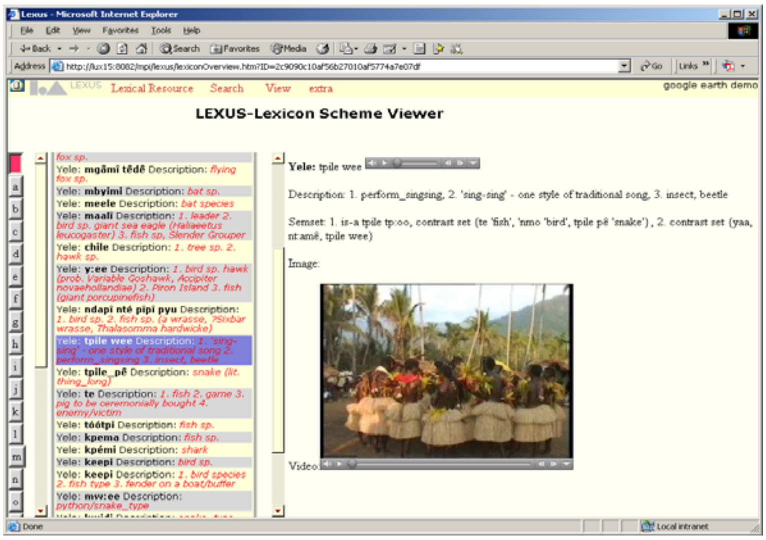
3.3 Summary

The two systems necessarily differ in their approach to the semantic integration process. While the MPI archive primarily relies on a lightweight process partly initiated by the collection creator, the TDS utilizes detailed metadata and semantic categories, whose integration requires considerable expert knowledge of the system. The semantic integration in the MPI case is basically left to the researcher who wants to discover interesting phenomena. In the TDS case the semantic integration is carried out by experts. Both approaches have their advantages and disadvantages: in the MPI case much work is left to the user, however, he can adapt his relations easily depending on the particular task. In the TDS case the user can immediately use the domain ontology, however, its definitions and relations are fixed. Due to the open design of the MPI framework, these two semantic domains can be integrated by taking over the concepts and relations from the TDS Global Ontaology into the personal concept and relation registries. Help needs to be given, since this integration step will require a deeper insight in the multi level ontology of TDS.

4 New Research Opportunities

Both initiatives focused on the development of web-based services that allow users to operate in an integrated domain of language resources via web-based interfaces. This has already been done for accessing the IMDI[15]

[15] IMDI provides a web-based metadata infrastructure, http://www.mpi.nl/imdi


Lexus - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
LEXUS
Lexical Resource Search View extra
google earth demo
a b c d e f g h i j k l m n o
Done
Local intranet

http://lux15:8082/mpi/lexus/lexiconOverview.htm?ID=2c9090c10af56b27010af5774a7e07df

LEXUS-Lexicon Scheme Viewer

fox sp.
Yele: **mgâmî têdê** Description: *flying fox sp.*
Yele: **mbyimi** Description: *bat sp.*
Yele: **meele** Description: *bat species*
Yele: **maali** Description: *1. leader 2. bird sp. giant sea eagle (Haliaeetus leucogaster) 3. fish sp, Slender Grouper*
Yele: **chile** Description: *1. tree sp. 2. hawk sp.*
Yele: **y:ee** Description: *1. bird sp. hawk (prob. Variable Goshawk, Accipiter novaehollandiae) 2. Piron Island 3. fish (giant porcupinefish)*
Yele: **ndapi nté pipi pyu** Description: *1. bird sp. 2. fish sp. (a wrasse, ?Sixbar wrasse, Thalasomma hardwicke)*
Yele: **tpile wee** Description: *1. 'sing-sing' - one style of traditional song 2. perform_singsing 3. insect, beetle*
Yele: **tpile_pê** Description: *snake (lit. thing_long)*
Yele: **te** Description: *1. fish 2. game 3. pig to be ceremonially bought 4. enemy/victim*
Yele: **tóótpi** Description: *fish sp.*
Yele: **kpema** Description: *fish sp.*
Yele: **kpémi** Description: *shark*
Yele: **keepi** Description: *bird sp.*
Yele: **keepi** Description: *1. bird species 2. fish type 3. fender on a boat/buffer*
Yele: **mw:ee** Description: *python/snake_type*

Yele: tpile wee

Description: 1. perform_singsing, 2. 'sing-sing' - one style of traditional song, 3. insect, beetle

Semset: 1. is-a tpile tp:oo, contrast set (te 'fish', 'nmo 'bird', tpile pê 'snake') , 2. contrast set (yaa, nt:amê, tpile wee)

Image:

Video:

Figure 4: Snapshot of the web-based LEXUS tool supporting the creation and manipluation of LMF-based lexica and also offering a web-service interface to access lexica.

metadata information, the annotations via ANNEX[16], the lexica via LEXUS[17], the TDS search engine and the central ISO reference Data Category Registry. Step-wised these will be transformed into real web-services with explicitly specified programming interfaces allowing developers to write new types of integrated applications such as

- searching across archives: search for certain patterns in all collections and databases by making use of the ontology frameworks
- operating across linguistic data types: look for patterns in a selection of annotations for a claim found in the typological database about certain languages and include them as references; present the complex lexicon entry for a word found in an annotation; carry out abstractions from a selection of resources and enter the results found into the typological database;
- web-based enrichment by commentaries: a framework is being developed that allows authorized persons to add comments to fragments of web-content presented by the various services mentioned; this includes to draw relations between such fragments

[16] ANNEX is a web-application provided by the MPI to access multimedia annotations, http://www.mpi.nl/lat

[17] LEXUS is a web-application provided by the MPI to access multimedia lexica, http://www.mpi.nl/lat

- web-based collaboration: the web-based services as described support intensive web-based collaboration when working in complex semantic domains; in particular the creation and management of shared "personal" ontologies will become an area where intensive collaboration will be required.

In particular the second option opens a vast number of new research opportunities in an integrated domain of linguistic resources and beyond that could be identified as the beginning of a true eHumanities domain, since they can be used by all researchers working with language material.

In figure 4 a snapshot of the user interface of the LEXUS web-application is shown as an example for the mentioned web-applications. Already now LEXUS offers a web-service with a standardized programming interface to access complex LMF-based lexica. It also interacts via web-services with the current ISO DCR implementation.

5 Conclusions

Currently all important components for the archive federation have been tested and we hope to soon bring substantial amounts of the partner's resources within the integrated virtual archive. The small number of participants in the DAM-LR project has enabled us to discuss and experiment with solutions for sensitive questions as authentication and authorization more efficiently then would have been possible with a larger number of participants. Within other projects we will be looking to enlarge the federation not only with other Language Resource archives but also with repositories of related disciplines in the humanities. Enlarging the federation will be possible thanks to the pioneering work done within DAM-LR but will also require to tackle issues that were less urgent with a small crowd like giving it a more stable "legal" basis. For the researchers institutional boundaries will become transparent.

Achieving semantic interoperability will require various approaches as could be shown and much handwork by experts and the users. In particular in open domains such as in large language resource archives it will be the user who needs to select the relevant concepts, to find out their different realizations in the various collections and create goal driven relations. Only very efficient frameworks supporting fast navigation, look-up, linking and re-using will help users to overcome the existing barriers. We foresee that much educational and training efforts will also be needed to convince researchers to make the required time investments. Finally, a broad acceptance with active contributions of many researchers is needed to create the network of shared ontologies that will drop the costs of cross-collection operations.

Compliant with the vision indicated by John Taylor, cited in the introduction, we have explained that for us the term "global collaboration" includes the notion of being able to access a large domain of virtually integrated resources seamlessly, efficient methods of overcoming the semantic differences with a variety of approaches and that a new "type of infrastructure" is required based on a Service Oriented Architecture offering many services that can be accessed via standardized interfaces.