An Evolving eScience Environment for Research Data in Linguistics

Claus Zinn, Peter Wittenburg, and Jacquelijn Ringersma

Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
firstname.lastname@mpi.nl

Abstract

The amount of research data in the Humanities is increasing at fast speed. Metadata helps describing and making accessible this data to interested researchers within and across institutions. While metadata interoperability is an issue that is being recognised and addressed, the systematic and user-driven provision of annotations and the linking together of resources into new organisational layers have received much less attention. This paper gives an overview of our evolving technological eScience environment to support such functionality. It describes two tools, ADDIT and ViCoS, which enable researchers, rather than archive managers, to organise and reorganise research data to fit their particular needs. The two tools, which are embedded into our institute's existing software landscape, are an initial step towards an eScience environment that gives our scientists easy access to (multimodal) research data of their interest, and empowers them to structure, enrich, link together, and share such data as they wish.

1. Introduction

The Max Planck Institute for Psycholinguistics (MPI-PL) holds more than 40 terabyte of research data in the area of language documentation. In the past 10 years, it has developed and deployed the Language Archiving Technology suite that helps researchers and archive management to digitize, ingest, catalogue, describe, and subsequently access all data. The sheer vastness of research data led to various tools for browsing and viewing research data, including the IMDI Browser and metadata search engine (Broeder et al., 2004), the ANNEX annotation exploration tool including the TROVA content search (Berck and Russel, 2006), geographic browsing (Uytvanck et al., 2008), and more recently, facetted browsing. Once research data is identified, however, users would want to further enrich it with commentaries and other types of annotations, or to link together resources via some semantic relation. In brief, they require means to organise and reorganise research data to fit their particular needs.

1.1. Managing research data at the MPI-PL

The archive of the MPI-PL hosts various forms of data in various media types. In the DoBeS project (Wittenburg, 2003) (Documentation of Endangered Languages, see `http://www.mpi.nl/dobes/`), rich resources have been gathered in the form of audio, video, and photographic material. All primary sources were catalogued with metadata using the IMDI editor. Most researchers describe their resources in terms of the content (*e.g.*, language, genre, modality); often researchers also add metadata to the actors (*e.g.*, interviewer, interviewee) together with actor details (*e.g.*, age or gender). Information on the quality of the resources is sometimes given, indicating whether the recordings were made under windy field or laboratory conditions. There is also some technical metadata of the resources (file type, format, size), which is determined automatically from the resource file during the upload process.

To a smaller but substantial extent, primary data is annotated along a multitude of dimensions using ELAN (Wittenburg et al., 2006). Often, for instance, alignments are being produced between transcriptions and audio signals, or annotation layers for speakers, voices, or voice qualities are being created. Another substantial type of resource created by MPI-PL researchers are lexica of the languages they document. Using MPI-PL's in-house tool LEXUS (Kemps-Snijders et al., 2006; Cablitz et al., 2007), researchers can define a lexicon's schema and then add entries as instantiations of the schema. LEXUS supports the inclusion of multimedia media either stored locally or from archived material. Photographic, audio and video material can thus be part of the lexicon to better illustrate the usage of a word in its social and cultural context.

1.2. New Requirements

Despite the variety of tools the amount of data can be overwhelming for users. Thus, users would profit from another layer of organisation where the space of research data can be organized to their needs, and where research data can be connected across the various types of data (metadata, multimedia data, lexicon data *etc.*). To achieve this, we aim at building a framework that allows researchers:

- to create their own user-defined workspace as selection and view of all archived material;
- to make accessible (a part of) their workspace to other researchers to foster collaboration;
- to attach notes, commentaries, to-do lists, *etc.* to each element (or their parts) of the workspace;
- to link together the various elements (or their parts) of the workspace by a variety of user-definable semantic relations.

We have built initial software towards achieving these aims, which we present next.

2. A First Step: ADDIT

A first step towards a conceptual (re-)organisation and enrichment of research data was taken with ADDIT, a web-based tool that allows users to add webnotes (or commentaries) to multimedia segments (video, audio, image data),

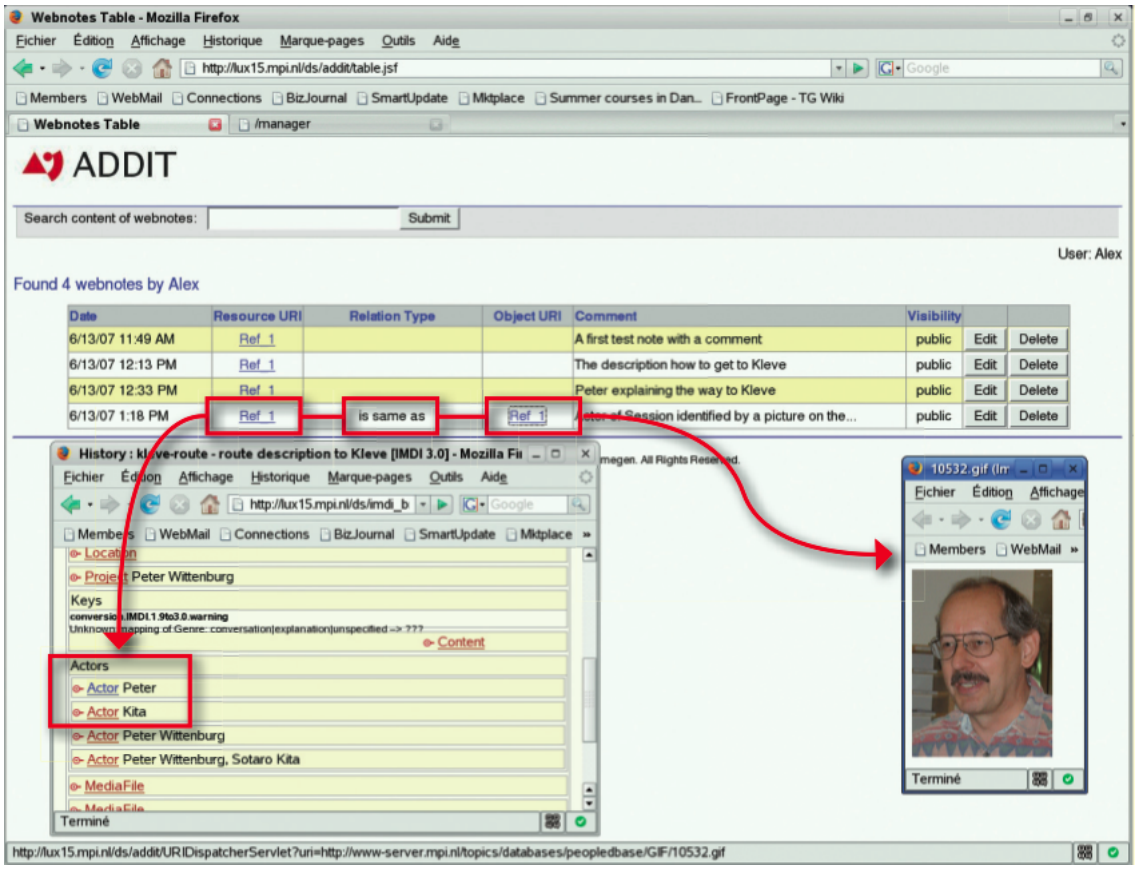


Figure 1: ADDIT Main Window

metadata records, other structured textual data such as annotations, and lexical data fragments. These full-text commentaries could direct researchers to review a particular multimedia resource, its metadata description, an annotation or a lexical entry. ADDIT also supports the linking together of research data of various types *via* semantic relation, for instance, to create genealogical relations between the interviewed persons of a large media resource (linking together parts of IMDI records), synonym relations (linking together entries of lexical resources), or connections across resource types, *e.g.*, linking together a lexeme from a lexicon with an audio fragment where it occurs.

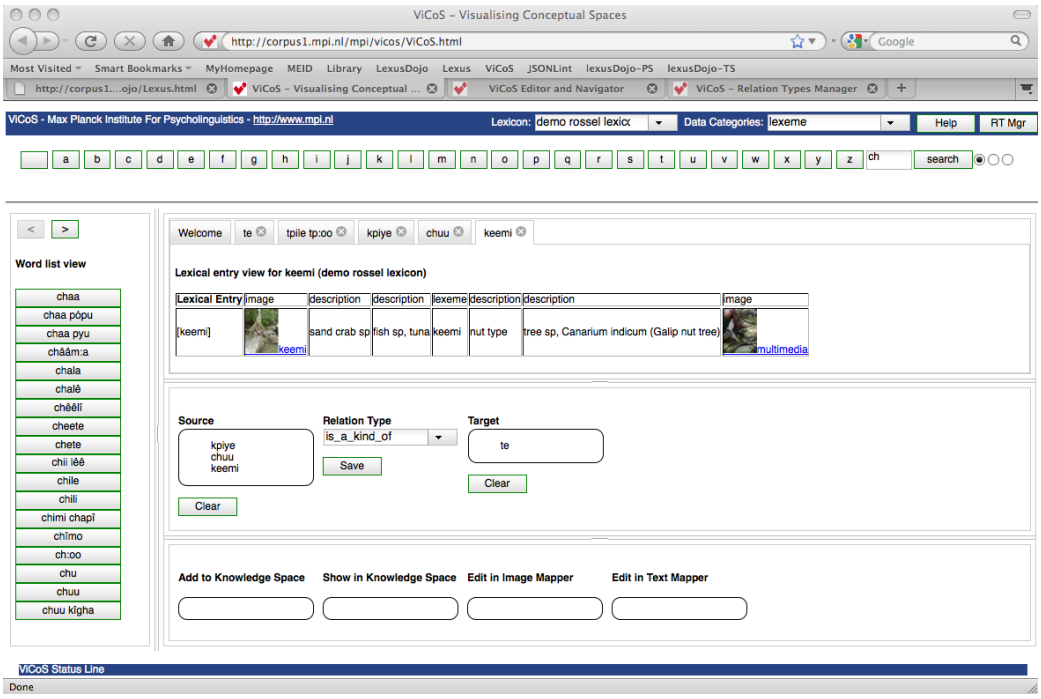
Fig. 1 shows ADDIT's main window with three webnotes and one semantic relation, which is detailed. Here, the metadata field "Actor" with value "Peter" is linked to an archived photo via a "is same as" relation. Every webnote or semantic relation created is associated with the one (webnote) or two URIs (semantic relations) it enriches, a comment to describe it, the type of relations (for semantic relations), and other values such as date of creation, creator, and access permissions.

ADDIT requires MPI-PL's web applications such as ANNEX, IMDI and LEXUS to adhere to ADDITs application programming interface (API) and to deliver unique resource or resource fragment identifiers, which could then be stored in ADDIT's data storage. ANNEX, IMDI and LEXUS can query ADDIT to find out whether there are webnotes or semantic relations defined for a given resource fragment, and if existing, to display such information to their users. ADDIT is implemented using Java with JavaServerFaces (see `http://java.sun.com/javaee/javaserverfaces/`), and the text engine Lucene (see `http://lucene.apache.org/`).
While ADDIT supports various LAT tools, its feature base is rather shallow and its user interface rather poor. With ViCoS, we aim at exploring a wide range of features, supported with a user-friendly and attractive user interface, but focus on enriching lexical resources only.

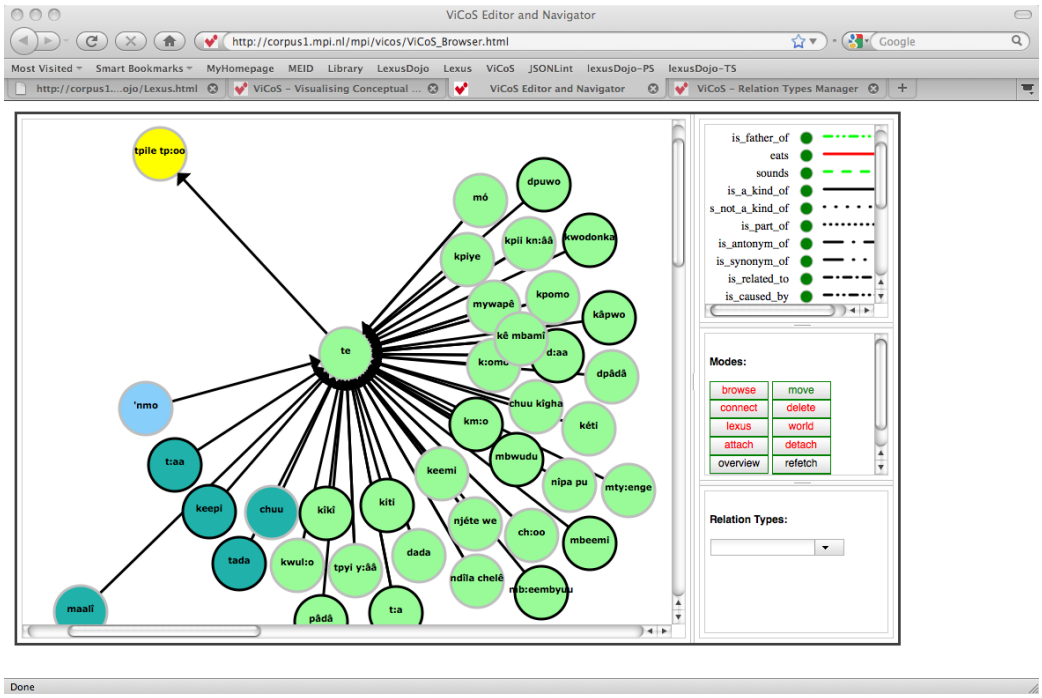
3. ViCoS

The current version of ViCoS is designed to add value to LEXUS, allowing users to add an additional layer to existing data, complementing a given linguistic description with a conceptual account. ViCoS' motivation in this setting is that a language is so much more than a list of lexical entries and their scientific description in linguistic parlour. With ViCoS, words can be turned into culturally relevant concepts and placed in relation to other concepts. Users can browse between lexical and ontological space more or less simultaneously and can thus gain a richer experience of the language and culture being documented.[1] Fig. 2(a) depicts ViCoS' main window, indicating how users can select

[1]The following research data stems from the Yélî Dnye project of Stephen Levinson's group, see the website at `http://www.mpi.nl/institute/research-groups/language-and-cognition-group/fieldsites/yeli-dnye`. Yélî Dnye is a language spoken on Rossel Island,



(a) ViCoS Main Window



(b) ViCoS Browser Window

Figure 2: The ViCoS Tool

lexical information and create semantic relations between them. ViCoS offers a few standard pre-defined "universal" relation types such as hyponomy, meronymy, homonymy, synonymy, and antonymy, but users can also freely define new relation types. Here, users are asked to label the relation type, give some informal description, choose a visualisation (line type and colour), and specify whether the relation type is functional, symmetric or transitive. To strive for semantic interoperability, we also foresee an interface to the relation types stored in the ISO Data Category Registry (Wright, 2000; Offenga et al., 2006) using the RESTful web services of the ISOcat software (see

Papua New Guinea, and MPI researchers have collected rich data sets encompassing many images, videos, and annotations thereof. There is also a lexicon of more than 6000 lexical entries, created with LEXUS, which also contains references to objects and species in the natural world, and links to corresponding material in the archive. In this context, ViCoS is being used to complement the lexical space with a conceptual space, to represent various cultural aspects of the language community and the natural world.

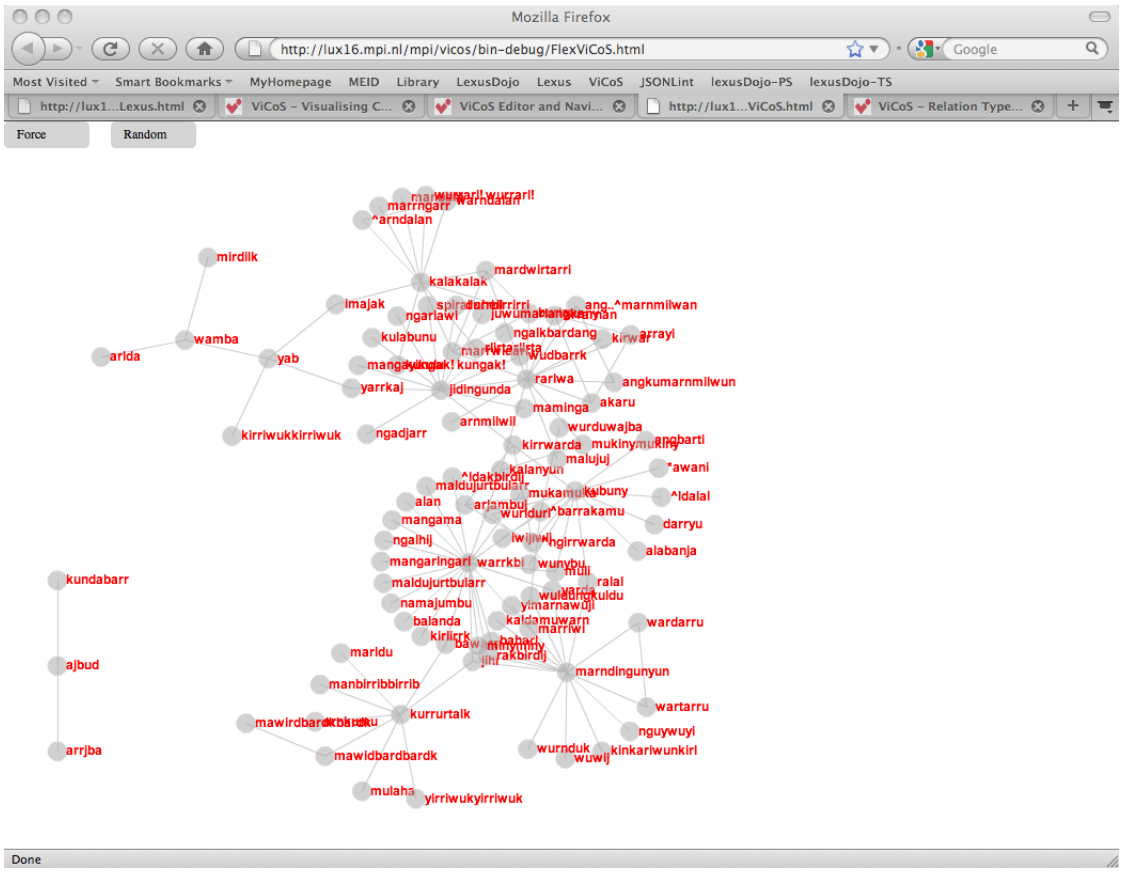


Figure 3: Overview functionality to display the complete conceptual space

`http://www.isocat.org`).

To view the conceptual space for a lexical entry, the user drags this lexical entry into the "Show in Knowledge Space" area. The lexical entry's corresponding concept will take center stage in the conceptual space browser, surrounded by all concepts that are *directly* related to it (see Fig. 2(b)).

The conceptual space browser offers, among others, the following modes (for more details, see (Zinn, 2008)). In *LEXUS mode*, clicking on the node will open LEXUS to show the corresponding lexical entry in linguistic context, thus strongly linking together lexical and ontological spaces. In *attach mode*, it is possible to link arbitrary URIs to a node; in "world" mode, clicking on a node with border will open the URI in a new browser tab. Users can thus link concept nodes to URI-accessible material of the MPI archive. If the URI, for instance, points to an ELAN file, the ANNEX viewer opens and shows the multimedia file and its annotations (playing the file from a time as retrieved from the time stamp parameter of the URI). In *colour mode*, colours can be attached to nodes; here colour coding is entirely up to the user. The depicted conceptual space, for instance, uses different colours to represent the generic term for bird and fish, and then other colours for bird and fish instances.

The ViCoS Browser Window only shows direct relations between the "center concept" and other concepts. To obtain a full overview, users can press the mode button "overview", and then the complete knowledge space is being displayed (see Fig. 3). The overview helps researchers to get a good first impression of his data; in our example, one can see clusters of bird, fish, crab, and man instances.

On the front-end, ViCoS uses the AJAX framework Dojo (see `http://www.dojotoolkit.org/`), which communicates with LEXUS *via* http requests. On the back-end, ViCoS uses Java-based code on top of the OWL knowledge representation language (McGuinness and van Harmelen, 2004). Conceptual spaces are stored within the JENA framework (see `http://jena.sourceforge.net/`), which provides a programmatic environment for OWL constructs and a rule-based inference engine for retrieving implicit information.

4. Discussion

So far, the current usage of ADDIT among MPI-PL researchers was rather disappointing. The reasons for this could be manyfold: there has been no systematic promotion of the tool within MPI-PL's researcher community, and thus only few commentaries or webnotes have been created so far. Moreover, ADDIT's UI is rather simple, and its capabilities to search for or display available webnotes and semantic relations is rather poor.

The contrary is rather true for ViCoS, which is building up an enthusiastic user community. A recent demonstration of ViCoS at a DoBeS training workshop showed a large interest of linguists in constructing and visualising conceptual spaces. The powerful but user-friendly UI was received

positively, and requests for new features come in regularly. A number of large conceptual spaces (each comprising at least hundreds of nodes) show a continuous and intensive use of ViCoS.

ViCoS has thus developed into a convincing showcase for enriching existing resources (in the current case, lexical resources) with additional information (such as colour coding), and for linking together lexical entries, or their parts, within or across lexica, or to arbitrary URIs. The latter makes it indeed possible to create links between a lexicon part and resource fragments that can be metadata, multimedia or annotations. However, in this respect, ViCoS is less powerful than ADDIT as the latter can also communicate and exchange data with tools for metadata management and multimedia annotation. Nevertheless, ViCoS is a good starting point to work towards fully-fledged tool support for linking together every piece of our research data with any other element inside or outside the institute's domain.

5. Future Work

In the following, we describe some of the use cases that we are currently pursuing to push further towards our vision of an integrated eScience environment.

5.1. Local Metadata Views.

Currently, all metadata-based access to archived resources is global, *i.e.*, all IMDI users see the same structured view. Moreover, all changes in the IMDI metadata tree require the involvement of archive management, and existing multiple paths to identical resources, while useful for some, may confuse others. In the spirit of ViCoS, we aim at enabling researchers to have their data organised themselves according to their individual – and potentially, dynamically changing – needs. We aim, thus, at enabling ViCoS to create links across elements of the IMDI metadata space; this would require an IMDI-ViCoS interaction protocol, similar to the one already in place for LEXUS-ViCoS.

Example Use Case. The MPI-PL archive hosts the Corpus for Dutch Sign Language, with 13.000 video and 2.500 ELAN files. A video file and its associated annotation file, if existing, are bundled together and described by IMDI metadata. While the whole set of video, annotation and IMDI file is referred to as *session resource bundle*, each of the three individual files are uniquely identified with an archive handle.

Currently, the Dutch Sign Language data is organized along three dimensions, namely, by recording, recording type (Canary Row, Fable Stories, Spot the Difference *etc.*), and region (Amsterdam, Rotterdam *etc.*). The same session resource bundle can be thus accessed through any of these three dimensions. This organisation is achieved without the generation of file copies in the archive; rather, session resource bundles are simply referred to from different parent nodes in the IMDI database.

While these metadata dimensions certainly respond to most researchers' needs, the definition of new organisational principles requires the involvement of archive management, and all changes in organisation are always global and accessible to all.

In the spirit of ViCoS, we aim at enabling researchers to have their data organised themselves according to their individual – and potentially, dynamically changing – needs. Reconsider the Dutch Sign Language corpus. Here, in total 92 signers have been recorded, and we assume that a researcher may want to *quickly* access all annotations for a given actor (or set of actors). In fact, an actor usually only appears in a small subset of the video files, and within a given video only in some of its segments. Fortunately, it is likely that the video's corresponding annotation file already has an ELAN tier to annotate speakers so that there is existing (time codings) information in which video segments a given speaker appears. Since the ELAN file has a unique identifier, there is thus a unique way to refer to each of the annotations of an actor in a given video: the file UID, the tier definition (or label), and a timecode for the beginning and end of the segment. With this information, it should be possible to link the appropriate IMDI metadata record "Actor" with its value to all occurences of this actor in the video resources, given that not only the IMDI file is uniquely identified, but also its metadata fields (in particular, the "Actor" field). Once each actor element in the IMDI file can be uniquely identified, a next version of ViCoS can be used to produce semantic links of the aforementioned type, but also other links to, say, relate actors and their characteristics with each other.

5.2. Multi-Directional Linking.

With LEXUS, users can build a lexical layer over archived research data, making thus a lexical resource a central entry point to all language-related multimedia material. Users can link, for instance, the lexical entry denoting a certain ritual song and dance to a video file in the archive depicting the song or dance is question. The reverse direction is also of interest, but not yet implemented. Here, we would like to support researchers in the process of annotating audio or video resources, in particular, with regards to transcription. When users create such an annotation tier in ELAN, they should be enabled to link tier elements (words or expressions) directly to lexical entries as maintained by LEXUS. As a result, researchers would get direct access to a linguistic description *via* the spoken word, and *vice versa*.

5.3. Personal Workspaces.

Our previous work and the aforementioned use cases already point into the direction of personal workspaces. While currently, we allow researchers to access archived material from lexical resources (using LEXUS), create additional layers on top of lexical spaces (using ViCoS), we now aim at allowing them to directly reorganise (mostly filter) global metadata trees into their own metadata-based views, and giving them the opportunity to strongly link annotation tiers (as created with ELAN) to lexical resources (maintained by LEXUS) or ontological resources (maintained by ViCoS).

With personal workspaces, users would be able to drag&drop resources, as indexed by the global IMDI metadata tree, into their own local area. Existing metadata descriptions of the local resource could be inherited from the global one and later redefined or extended. Once a

workspace is populated, we would like to allow researchers to create links between its inhabitants, thus inducing a new organisational structure, or to attach notes to them.
Also, each resource will have its file type so that a double click on a local resource of the workspace would open the corresponding application to browse or edit metadata (IMDI), lexical resources (LEXUS), multimedia resources and annotation files (ELAN), and ontological resources (ViCoS). Once such an application is opened, it will offer access to its file content, but also across resources – as it is already possible with LEXUS (access to archive material), ViCoS (access to lexical resources), and as we envision it for ELAN (access to lexical resources) and IMDI (access to a particular audio/video segment plus tier).
In addition, once an application is open, users should be able to mark any resource fragment and also drag&drop the selection from inside the application to the outside, the users' local workspace. Here, the new workspace element can then be subject to further semantic enrichment, so that notes can be attached to it, and where it can be linked to other elements in the workspace.
Clearly, we could elaborate on the notion of the local workspace, for instance, in terms of sharing it (or parts thereof) across members of a research team to facilitate and foster collaboration.

6. Conclusion

Building an eScience environment to support researchers managing, enriching and exploiting their research data largely depends on an institution's existing tools, and the way they are used. Clearly, existing tool use and workflows should be taken into account, and interoperability between all tools is desired. Consequently, our description is quite specific to our existing research environment.
There are, however, quite a few general lessons learned. Where research data is archived, its elements shall be accessible *via* persistent identifiers; where element fragments become the object of enrichment, or the source or target of a semantic relation, appropriate and unique fragment paths need to be added to such identifiers. Another important issue is synchronisation support. When a workspace gives a researcher-specific view of his or her research data, the issue has to be addressed how to deal with situations where underlying research data is deleted or modified, potentially destroying some of the connections constructed in the workspace.
Our institute has put into place a persistent identifier scheme based upon the HANDLE system (see `http://www.handle.net`). Each archived resource can thus be uniquely and persistenly accessed. The situation with persistent access to fragments of a resource is trickier. LEXUS, for instance, maintains its own local database to identify lexicon names, their schemas and lexical entries, and their parts. Similarly, ELAN has its internal representation of tier definitions, and the individuals labels that occur on tiers. Also, there is a need to define how to address metadata fragments from the IMDI Browser. Here, a coordinated approach across the corresponding application designers is needed, and substantial code refactoring at the tools' back-ends. Also, the tools' user interfaces need to be extended to support the drag&drop operations we envision to move resource fragments across applications and the users' local workspace. Our initial expertise with the ADDIT and ViCoS software as well as the overwhelmingly positive feedback from ViCoS users, however, demonstrate the benefits and the feasability of the work ahead, towards a truly effective eScience research environment at our institute.

7. References

P. Berck and A. Russel. 2006. ANNEX - a web-based framework for exploiting annotated media resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

D. Broeder, T. Declerck, L. Romary, M. Uneson, S. Strömqvist, and P. Wittenburg. 2004. A large metadata domain of language resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

G. Cablitz, J. Ringersma, and M. Kemps-Snijders. 2007. Visualizing endangered indigenous languages of French Polynesia with LEXUS. In *Proceedings of the 11th International Conference Information Visualization (IV07)*, pages 409–414. IEEE Computer Society.

M. Kemps-Snijders, M-J. Nederhof, and P. Wittenburg. 2006. LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

D. L. McGuinness and F. van Harmelen. 2004. OWL web ontology language overview. w3c recommendation from 10 february 2004. Available at http://www.w3.org/TR/owl-features/.

F. Offenga, D. Broeder, P. Wittenburg, J. Ducret, and L. Romary. 2006. Metadata profile in the ISO data category registry. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

D. Van Uytvanck, A. Dukers, J. Ringersma, and P. Trilsbeek. 2008. Language-sites: Accessing and presenting language resources via geographic information systems. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

P. Wittenburg. 2003. The DoBeS model of language documentation. *Language Documentation and Description*, 1:122–140.

S. E. Wright. 2000. A global Data Category Registry for interoperable language resources. Available at http://isotc.iso.org (ISO TC 37).

C. Zinn. 2008. Conceptual spaces in ViCoS. In Sean Bechhofer et al., editor, *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, volume LNCS 5021, pages 890–894, Tenerife. Springer-Verlag. Demo Track.