Exploring and Enriching a Language Resource Archive via the Web

M. Kemps-Snijders, A. Klassmann, C. Zinn, P. Berck, A. Russel, P. Wittenburg

Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Marc.Kemps-Snijders@mpi.nl , Alex.Klassmann@mpi.nl , Claus.Zinn@mpi.nl
Peter.Berck@mpi.nl, Albert.Russel@mpi.nl , Peter.Wittenburg@mpi.nl

Abstract

The "download first, then process paradigm" is still the predominant working method amongst the research community. The web-based paradigm, however, offers many advantages from a tool development and data management perspective as they allow a quick adaptation to changing research environments. Moreover, new ways of combining tools and data are increasingly becoming available and will eventually enable a true web-based workflow approach, thus challenging the "download first, then process" paradigm. The necessary infrastructure for managing, exploring and enriching language resources via the Web will need to be delivered by projects like CLARIN and DARIAH.

1. Introduction

Most researchers making use of language resources are still following the "download first, then process" paradigm: first, data needed to tackle a research question is collected and stored to their local computers, and only then data is processed and analyzed. This paradigm is most useful in research scenarios where, for instance,

- performance considerations require fast access to large amounts of data (i.e., identifying the parameters of a stochastic models to train speech recognizers);
- there is only limited, or low band-width Internet connection (e.g., field work in remote areas);
- data annotation requires high accuracy or precise timing, which is currently not supported by Internet protocols for data exchange (e.g., transcription of media streams); and
- where language technology tools are not yet available for online use.

In the long-term, however, language resources and tools will follow a web-based paradigm, where application service providers will offer access to resources via Internet based services. One popular example is Google Apps `[www.google.com/a]`, which already delivers web-based services around text processing, spread sheets, and calendar utilities to an increasing user base. Here, data is stored in remote storage systems and users can readily access their data via web applications from any place that offers Internet access and an Internet browser.

The web-based paradigm has many advantages over the traditional approach, and this paper will discuss them from the perspective of managing language resources and tools development. Moreover, the web-based paradigm opens new opportunities for sharing, enriching and linking together resources, which the paper briefly sketches.

2. Advantages of Web-based Paradigm

In the web-based paradigm, tools do not longer need to be installed and maintained on users' local computers. From a tool development and data management point of view this makes it easier to adapt to changing requirements and to propagate changes directly to users. This is a huge advantage to the traditional approach where the distribution of tools is done via software updates that requires user initiative and interaction; inevitably, this causes time delays between release date and prevents a rapid and wide-spread adoption among a broad user base. Moreover, in a web-based setting, the interaction between the development of tools and the design of data formats can be better synchronized. Tool builders and resource service centers can now interact directly to ensure that changes to resource formats are immediately propagated to the supporting tools, and conversely, extended tools functionality is rapidly reflected in resource formats. This facilitates data migration along software updates and makes it easier to ensure backward compatibility to legacy formats, thus allowing organizations involved in archiving activities to carefully orchestrate the evolution of resources and tools.

The web-based paradigm makes it possible to deliver new forms of technology that go beyond the sharing of resources and the management of data migration and software updates. Users may now use functionality that allows them, for instance,

- to create and follow-up references from an electronic publication to linguistic resources such as lexical entries, annotated media fragments or other electronic documents that can serve to support the claims made;
- to perform cross-walks between different linguistic data types, say by browsing a web-accessible lexicon while working on text annotations and *vice versa*;
- to bring together various types of resources from different disciplines, say via the creation of geographic overlay techniques, and in doing so, to gain new insights;

- to create virtual collections, potentially covering resources from various archives, to define temporary workspaces where researchers collaborate to tackle research questions at hand;

- to give commentaries to all sorts of resources and resource fragments, and to draw typed relations between such fragments; and

- to create views on collections that are different from, and for some users more attractive than, the canonical view defined by metadata; these views could be, for example, a knowledge space covering relevant concepts of a domain or a culture and their relations, with direct references from these concepts to all sorts of resource fragments.

These examples require interaction between multiple tools, where each tool contributes its specific functionality (such as retrieval and display of specific formats) to jointly process various aspects of the execution chain.

3. LAT Technology

The web-based paradigm will enable advanced cyber-infrastructure type of applications. The LAT technology developed at the Max Planck Institute for Psycholinguistics (MPI) is providing increasingly rich functionality along these lines. By including a reference in an electronic publication it is possible to invoke LEXUS (Kemps-Snijders et al., 2006) to visualize a lexicon fragment, or ANNEX (Berck and Russel, 2006) to visualize an annotated media fragment. An ISO standard proposal on such referencing in electronic resources has been drafted in collaboration with the DELAMAN network [www.delaman.org] of endangered languages and music archives and submitted to ISO TC37/SC4 [www.tc37sc4.org]. By using the same technique, crosswalks are possible between lexical resources and annotation resources; users can now easily take the headword of a lexicon and search for all its occurrences in the transcription, or while doing an annotation ask the lexical resource to suggest word completions.

3.1. Easy Access

Geographic systems are well suited for bringing together various types of diverse multi-disciplinary information. Language-Sites [www.mpi.nl/services/mpi-archive/GE_language_sites], also stemming from the DELAMAN network, informs users about existing language resources whenever zooming into a geographic region where such resources are available. However, users should not only obtain information about the availability of such resources, but also be enabled to directly access and manipulate them. The MPI investigated and deployed various possibilities such as immediately displaying the appropriate node in the archive, or starting the annotation of media streams or multimedia lexica that are stored in the archive. In a web-based framework, the realization of such behaviour is simple, using URLs pointing to web-applications and the appropriate resources.

3.2. The Dynamics of Resources

In the study of minority and endangered languages, there is an ever increasing set of data, and usually primary data (*e.g.*, multimedia recordings) outnumber secondary data (*e.g.*, metadata, annotations). Existing corpora become richer, and new corpora are being created by the day. Even large national corpora, such as the Dutch Spoken Language Corpus [http://www.tst.inl.nl/cgndocs/doc_English/start.htm], will change over time, with new resources being added to make them more balanced, to cover new varieties, or to consider language variations over time. The traditional static notion of corpora barely reflects the dynamics of such resources. Only the web-based paradigm with its distributed responsibilities will allow us to manage such corpora and to access them adequately. The static view of distributing "complete" corpora will need to be replaced with a dynamic one, where corpora can be accessed, joined, and analysed online. The IMDI metadata infrastructure (Broeder et al., 2004) together with the LAMUS archiving system (Broeder et al., 2006) is supporting these working methods.

3.3. Virtual Collections

Working on virtual collections that contain resources from different repositories is another dimension that has been addressed. Within the EC-funded DAM-LR project [www.mpi.nl/DAM-LR] middleware technologies such as the Handle System [www.handle.net] and Shibboleth [shibboleth.internet2.edu] have been installed at three archives that allow users to form integrated virtual collections based on integrated metadata domains. For this, web applications such as LEXUS and ANNEX need to be adapted to support workflows that cross repository boundaries. In particular, they need to support the challenges of distributed authentication and authorization, but also a common terminology (say, for effectively carrying out search across collections). For now, access management is in its preliminary stages, but there is considerable progress on terminology support. The emergence of the ISOcat Data Category Registry (DCR) (Offenga et al., 2006) is one cornerstone to tackle the problems of semantic interoperability, at least at the level of linguistic terminology. Within the LIRICS project [lirics.loria.fr] a standard API has been defined and implemented for the DCR. This is now also being implemented in the new ISOcat DCR software, running as a true web service. Tools such as LEXUS have been adapted to support the ISOcat API.

3.4. Commenting and Relating

Increasingly important for future research is to offer experts means to add commentaries to existing content and to draw semantic relations between resources or fragments thereof. Commentaries improve collaboration within groups, can serve to create folksonomies, and can help to improve and enrich web-accessible content. Semantic relations, whether created manually, semi-automatically or automatically, will help researchers to exploit rich content in many ways even in a cross-disciplinary research setting. With ADDIT [http://www.lat-mpi.eu/tools/

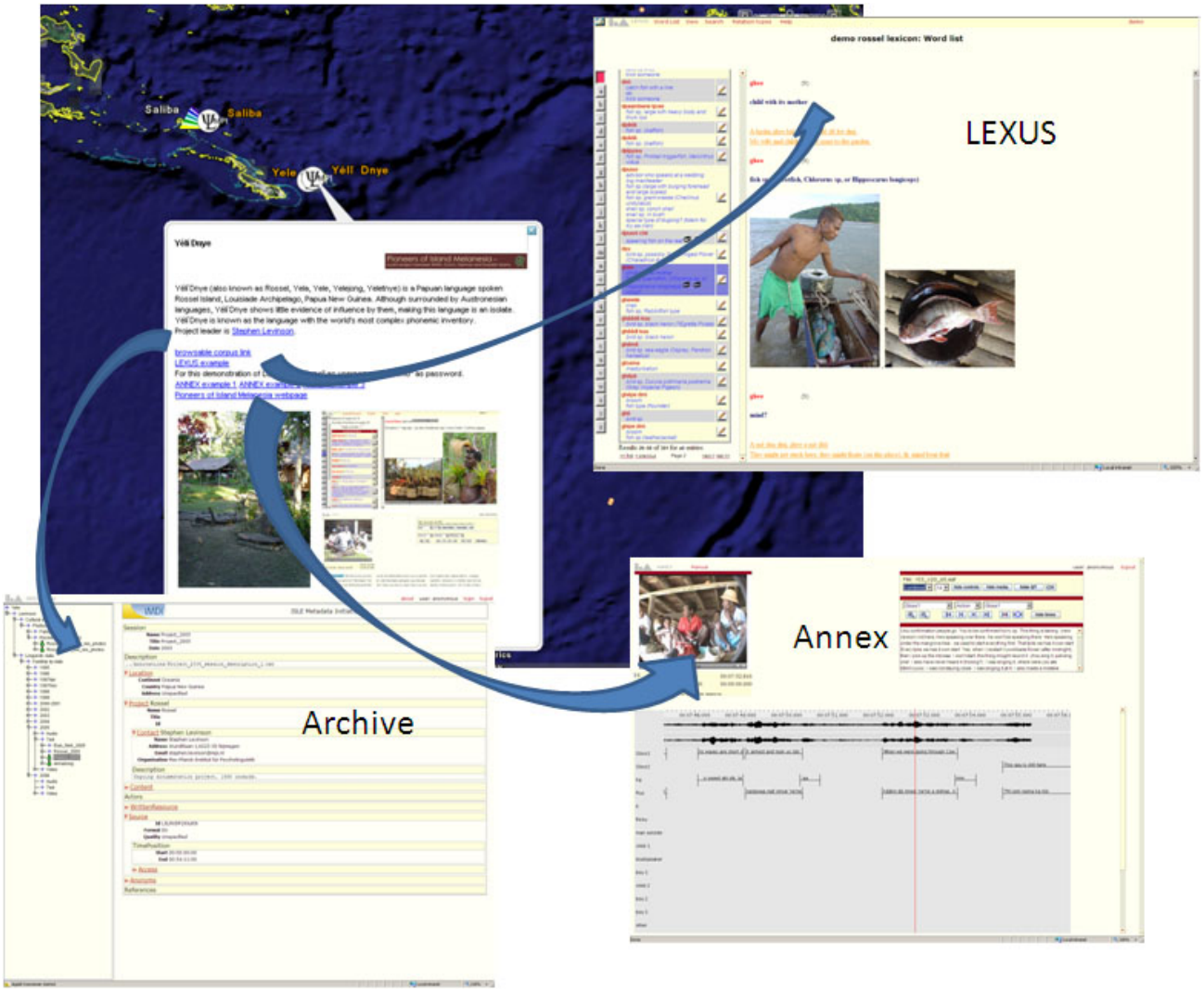

Figure 1: Google Earth overlay providing entry points to linguistic resources, see [www.mpi.nl/services/mpi-archive/GE_language_sites].

addit] and ViCoS (Zinn et al., 2008) the MPI has developed two tools that allow users: the creation of commentaries for content such as metadata, annotated media, multimedia lexica and arbitrary websites (as long as they provide static content); the definition of relations between such content; and the visualization and navigation of such content in the resulting "Conceptual Spaces". For the referencing of resource fragments, tools such as IMDI, LEXUS and ANNEX were adapted to provide access points; they will also inform users whenever comments or enrichments are available, given the entries of the databases for commentaries and conceptual spaces.

3.5. Community-specific Views

For some communities it is important to have a specific view on their material. This is particularly true for linguists and language communities in documentation or sign language projects, given that browsing and searching through metadata is considered neither intuitive nor attractive. Here, specially designed community websites can present resources in a much more focused or targeted view. This creates a challenge for modern language resources archives as they store and make available resources from many different communities in a more or less uniform way. To make feasible the effort of maintaining community portals, the MPI designed an initial set of web-templates that also embed simple command options, for example, for the execution of metadata searches in real time. This allows users to group, for instance, resources according to their genre and to offer them at the web-site in an actual state without the need to manually and continuously adapt links. The technology requires the availability of REST-based APIs to access resource indices and high quality metadata.

4. Example Use Case

Yélî Dnye is a Papuan language spoken on Rossel Island, Louisiade Archipelago, Papua New Guinea. The Yélî Dnye corpus currently consists of data from 1995 to present, including audio, video, images, texts and annotated media. Fig. 1 shows an appealing and popular way to select and enter this (and other) linguistic resources. Language Sites provides access to the corpus by means of geo-

graphical browsing, using a Google Earth overlay. The information associated with the overlay section provides a direct link to the corpus in the MPI Archive, a link to the Yélî Dnye lexicon and some sample links to annotated media files. Each of these links will invoke appropriate tools for displaying information in a user-oriented manner such as IMDI Browser, LEXUS or ANNEX.

Each of these tools may in turn use the same technique to display information handled by other tools. Both the IMDI Browser and LEXUS, for instance, will use ANNEX to display annotated multimedia files stored in the archive or as part of a lexical entry respectively. More complex interaction patterns that go beyond simple information display are also possible. With ViCoS, we are building a conceptual space around the Yélî Dnye lexicon, aiming at representing the natural world on Rossel island from an ethnobiological perspective. The ViCoS tool gains access to the lexicon via a LEXUS web service so that users can easily create semantic relations between concepts denoted by lexical entries or parts thereof. By browsing the conceptual map in ViCoS, users are only one click away from the corresponding lexical entries (and their display in LEXUS). On the other hand, LEXUS' display of lexical entries shows the availability of ViCoS content, which users can enter again by a 1-click mouse action. With ADDIT, users can add commentaries to and relations between content ("webnotes") maintained by ANNEX, LEXUS, and the IMDI Browser (and other ADDIT-compliant tools). With ADDIT, our experts for Yélî Dnye are empowered to, say, point at controversial elements in a linguistic or anthropological analysis, or to enrich a documentation by a comment, or to set personal reminders, or to support collaborative work on resources. Again, the availability of such commentary is highlighted by all tools that currently support the ADDIT API.

The technique of incorporating tools using the web service paradigm is not only limited to tools developed internally by the MPI, but also includes interaction with tools from third parties. LEXUS as well as the multimedia annotation tool ELAN (Wittenburg et al., 2006), for instance, interface with the Data Category Registry web service.

5. Conclusion

The MPI has adopted the web-based paradigm and will support an increasing number of web applications and services for accessing and manipulating linguistic resources. Those services, in turn, can be embedded by other services and applications by third party providers. Once a crucial set of services is available, web-based workflows will be a real alternative to the currently dominating "download first, then process" paradigm. In this paper, we have sketched some advantages and applications for the web-based workflow paradigm, and there will be many others that we cannot foresee yet. We need to consider, however, that for now, and for specific applications and circumstances, the "download first, then process" paradigm will remain the primary choice for many researchers. When the web-based paradigm joins forces with high performance grid-based solutions, the number of such circumstances and applications will certainly decrease and disappear as the main working method.

Although some institutes have already understood the potential of the new metaphor and are creating an increasing amount of services of all types, we miss the infrastructure and the critical mass to achieve a breakthrough. It will be the task of infrastructure projects such as CLARIN `[www.clarin.eu]` and DARIAH `[www.dariah.eu]` to come to such a critical mass of services and to achieve a broad coverage so that a large user base will adopt and benefit from the web-based workflow paradigm.

Links to web-based LAT Software

LEXUS	http://www.lat-mpi.eu/tools/lexus
ANNEX	http://www.lat-mpi.eu/tools/annex
ADDIT	http://www.lat-mpi.eu/tools/addit
VICOS	http://www.lat-mpi.eu/tools/vicos
LAMUS	http://www.lat-mpi.eu/tools/lamus
IMDI	http://www.mpi.nl/IMDI

6. 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)*.

D. Broeder, A. Claus, F. Offenga, R. Skiba, P. Trilsbeek, and P Wittenburg. 2006. Lamus: The language archive management and upload system. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2006. Lexical markup framework: Iso standard for semantic information in nlp lexicons. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Also, see http://www.lexicalmarkupframework.org.

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)*.

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)*.

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)*.

C. Zinn, G. Cablitz, J. Ringersma, M. Kemps-Snijders, and P. Wittenburg. 2008. Constructing knowledge spaces from linguistic resources. In *CIL 18 Workshop on Linguistic Studies of Ontology: From Lexical Semantics to Formal Ontologies and Back*.