Eureka! User friendly access to the MPI linguistic data archive

Jacquelijn Ringersma, Claus Zinn and Alexander Koenig
Max Planck Institute for Psycholinguistics (MPI)
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
{firstname.lastname@mpi.nl}

Abstract
The MPI archive hosts a rich and diverse set of linguistic resources, containing some 300.000 audio, video and text resources, which are described by some 100.000 metadata files. New data is ingested on a daily basis, and there is an increasing need to facilitate easy access to both expert and novice users. In this paper, we describe various tools that help users to view all archived content: the IMDI Browser, providing metadata-based access through structured tree navigation and search; a facetted browser where users select from a few distinctive metadata fields (facets) to find the resource(s) in need; a Google Earth overlay where resources can be located via geographic reference; purpose-built web portals giving pre-fabricated access to a well-defined part of the archive; lexicon-based entry points to parts of the archive where browsing a lexicon gives access to non-linguistic material; and finally, an ontology-based approach where lexical spaces are complemented with conceptual ones to give a more structured extra-linguistic view of the languages and cultures its helps documenting.

1. The MPI linguistic data archive

The Max Planck Institute for Psycholinguistics (MPI) conducts research on language comprehension, language acquisition, the relation between language, culture & cognition and on the neuro-biological basis of language. Around 10 years ago, a digital archive has been set up for the long term preservation and availability of research data. At the time of writing, the archive contains around 400.000 archived objects with a total size of around 16.5 Terabytes [1]. Besides the research data collected by MPI researchers, the archive also holds other data such as the catalogues from the Documentation of Endangered Languages (DoBeS) project [2]; within this project some 50 languages, which are threatened to become extinct in the next decade, are being documented through audio and video recordings and rich annotations of those. Moreover, the archive holds, among others: the *Corpus Nederlandse Gebarentaal* (NGT) [3], a unique sign language corpus of video data, totaling more than 70 hours of rich video material of 92 Dutch deaf people; the *Dutch Bilingual Database*, comprising some 2000 video, audio and text resources on (immigrant) bilingualism in the Netherlands [4]; and the European Science Foundation *Second Language Acquisition by Adult Immigrants* (ESF) corpus [5], which contains spontaneous second language acquisition data. A number of other institutions also make use of the MPI archive infrastructure: Leiden University, for instance, stores its Carib data collected in Surinam in the 1960's [6] and the University of Surrey stores its Slavonic colour lexicon [7].

A substantial amount of data in the archive is of major importance for research and education, and gives rich testament to human language and culture. According to

the UNESCO, about half of the 6,700 languages currently spoken is threatened to become extinct by the end of the 21st century [8], which gives the data in the DoBeS corpus particular importance. This data serves not only research, documentation and conservation purposes, but is also meant to play a major role for revitalization efforts of the languages and cultures in question. Archived data also plays a supporting role for education, such as the corpus on second language acquisition of immigrants. Another example is the NGT corpus, which is already used for the training of sign interpreters, and which, in addition, is contributing to the emancipation of deaf people in the Netherlands.

The archive's central organization concept is the *session resource bundle*. It bundles together all resources pertaining to the same linguistic event, usually comprising a set of media resources (audio, video, image), a set of written resources (annotation and transcription files), and a metadata file that describes all content. The metadata file follows the IMDI metadata scheme, which was developed especially for the description of linguistic resources [9]. The scheme contains metadata elements on the context in which the data was collected (project, contact person, location), elements which describe the content of the resources (language, genre, modality) and elements describing the resources in a more technical manner (format, size, resource link). Session resource bundles usually are grouped into (sub-) corpora.

In order to obtain persistency over time and independency of the resource location each archived object is assigned a unique resource identifier (URID) to provide a stable method for referencing electronic resources. Conceptually, URIDs can be compared with ISBN numbers that are used to uniquely identify published books. The MPI archive chose the widely used Handle System (HS) to create, maintain and resolve URIDs [10].

The archive is available online: however for most researchers and members of the speech communities it is a requirement that access to the resources in the archive can be restricted. For this, researchers can use our Access Management System to assign one of the following three levels of user access rights to each resource: access to all; access to registered users who have signed a Code of Conduct; and access to a selected group of users [11].

Overview. In the remainder of this paper, we describe different ways of accessing the archived data, accommodating the varying technical skills and needs of different stakeholders. Section 2 describes the traditional access method, the IMDI browser and search engine. It provides fully-fledged metadata-based access to all data, but requires significant knowledge of the archive's organization, content and the IMDI metadata fields. It is primarily targeted at the research community, in particular scientists who regularly contribute and exploit archived data. For members of the general public, with lesser familiarity with the archive, we provide two alternative methods of archive access: by facetted browsing (see Section 3) and by geographical browsing (Section 4). Members of the speech communities might have special interest only in the data to which they relate most. For this purpose we have developed special web portals which provide access to the data of one specific language, organized in a thematic manner with

a user interface that 'connects' to the cultural environment of the language in question. The web portals are described in Section 5. Finally, in Section 6 and 7, we describe LEXUS and ViCoS, two tightly coupled applications that allow users to create lexical and conceptual spaces through which the data in the archive can be accessed either by words (LEXUS) or by conceptual spaces anchored in the words of the lexicon (ViCoS). While LEXUS is a tool primarily targeted at linguists, ViCOS aims at attracting members of the speech communities as data can be accessed via networks of culturally relevant concepts, unobscured by linguistic detail and parlour.

2. IMDI Browser and search engine

The IMDI browser provides online access to the corpora stored in the MPI archive. The browser view is based on a database derived from the IMDI metadata in the session resource bundle's metadata files. For the creation of these files we developed the IMDI metadata editor [12], which provides a structured and easy manner to describe the resources with a full or limited set of the IMDI metadata.

The data in the browser is presented in a tree structure; directly visible top nodes represent the major corpora. The internal organization of a corpus is designed by the researcher. A corpus, for instance, can be organized by geographic location or by thematic subject or by a combination of the two. Figure 1 (left) shows the organization of the Marquesan corpus in the DoBeS archive. While the first two levels in the tree structure are organized by geographic location, the next level is organised by thematic subject. The session resource bundle (the green bag icon in Figure 1) is attached to the thematic corpus nodes.

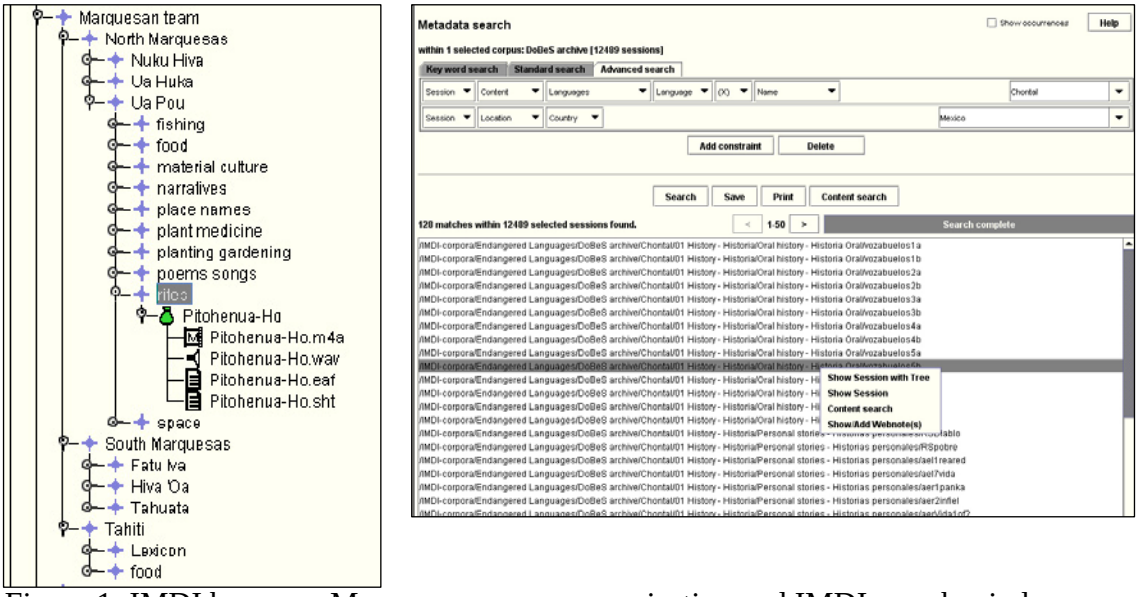

Figure 1: IMDI browser: Marquesan corpus organisation and IMDI search window

Visitors can browse the tree by clicking and opening the corpus nodes. The metadata describing the resources can be viewed directly from the tree browser, since metadata

never has any access restrictions. Finally the resource files can be viewed or downloaded from the browser, provided that the visitor has the required access rights.

Browsing the archive in this manner suits those researchers best who have a good knowledge of its structure and content. Without such knowledge it can be hard, for instance, to find all resources on 'fermented breadfruit food preparation' in the Marquesan language. Metadata-based browsing is thus complemented by metadata-based search, allowing users to perform a keyword search or more specific searches (see Figure 1, right) in the IMDI elements of the metadata file in the resource bundles.

3. Facetted browsing

The facetted browser provides user-friendly access to all data of the MPI archive via well defined elements taken from the IMDI metadata set (*facets).* With such a browser users can easily navigate through all data by selecting one or more facets, each of which is defining a certain view and subset of the data. Facets can be selected in any order, and under the hood, set intersection takes place to compute the number of data satisfying the given selections. For illustration, Figure 2 shows the main page of the web-based browser with the main facets for 'corpus', 'continent', 'country', 'language', 'organization', 'genre', 'interactivity', 'planning type', 'involvement', and 'social context'. The 'corpus' facet gives access to the various corpora in the archive (the numbers in parentheses show the number of objects in each corpus). When a user selects the 'dobes' corpus (17498 entries), all data with regard to the other facets is updated. Moreover, it is possible to perform a full-text search on all data, or on the subsets resulting from facet selections.

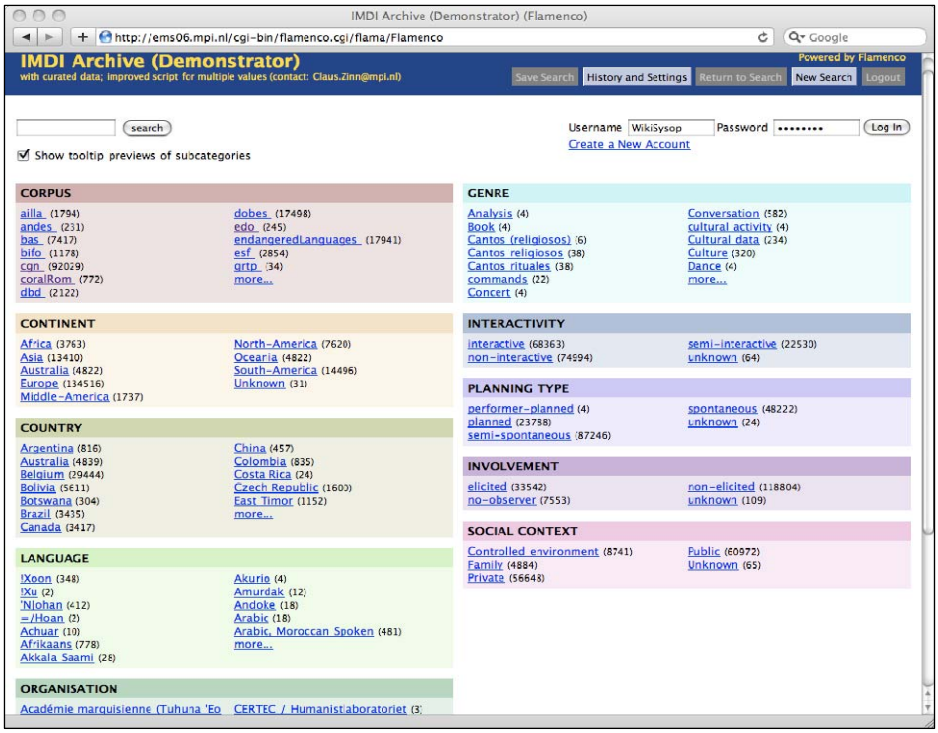

Figure 2: Overview of the MPIP archive through the facetted browser

Once a resource is identified, its full metadata is displayed, also indicating all the facets where it appears in. Here, users can find similar resources by further restricting their facet selection. Initial feedback from expert users is very positive. We expect novice users to profit from the facet browser as there is no need to learn a controlled language, or to have any other deep insight into archive organization and content. Moreover, the use of facetted browsing on the Internet is already quite common. For the time being, the facetted browser has demonstrator status, and more work is needed to increase response time – the 200.000+ entries of the MPI archive, with its 10 facets where each facet can have more than 300 different values, puts quite some strain on mySQL to compute the many set intersections. The demonstrator is using the *flamenco* facetted browser implementation from Berkeley [13], but we are also investigating whether Apache's *solr* framework [14], an actively maintained software, yields significant speed-ups.

4. Geographic navigation

With geographic navigation, we assume that novice users will search language resources by geographic area and that they are at ease with navigating through maps. For these users we provide a Google Earth (GE) [15] overlay for the integrated presentation and sharing of archived resources. GE was chosen because it is freely available for various operating systems, it has good navigation controls, and because the overlaid linguistic information can be stored in XML-based KML files, which facilitates its interchange with other geographic presentation frameworks. Last but not least, many users are familiar with the GE application. The MPI overlay includes references to archives, catalogues, resource bundles and anthropological sites [16, 17].

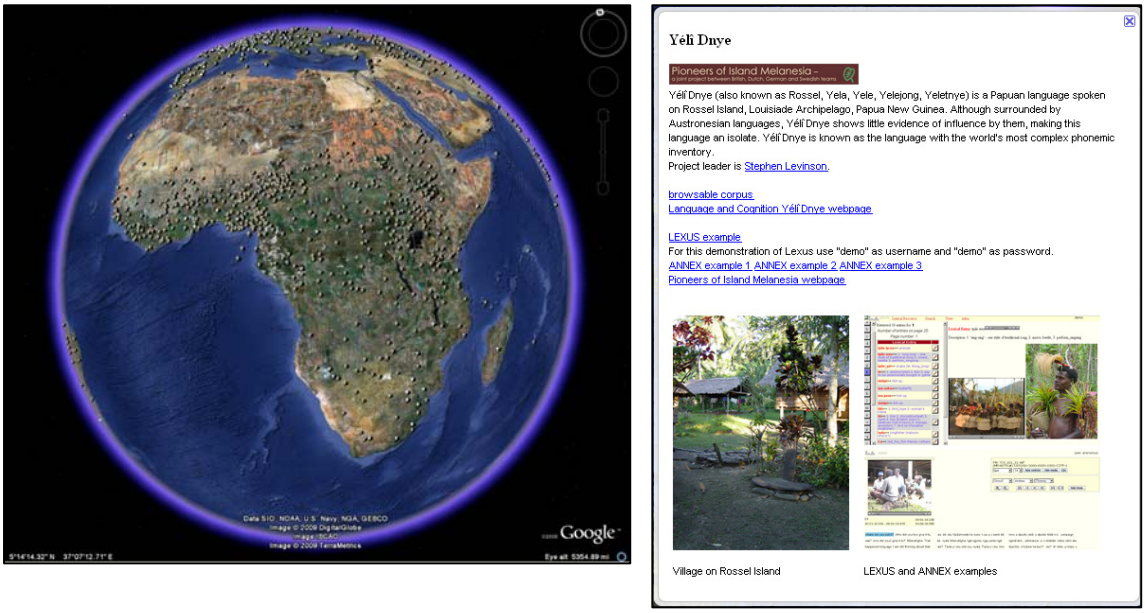

Figure 3: Overview of the GE overlay and pop-up for Yélî Dnye language resource

The GE overlay provides placemarks to language archives, not only from the MPI but also from other organizations storing language resources, such as Paradisec [18] or

AILLA [19]. A typical place mark is accompanied by a reference to the language archive and some introductory information. Besides pointers to archives and language catalogues, the GE overlay also provides placemarks as an entry point for (sets of) resource bundles. The text balloon that pops up when users click the landmark on the map can be enriched with introductory text, images, sound files and video. Persistent hyperlinks (handles) can be added in order to connect the places on the map with the associated resources in the MPI archive or elsewhere. Figure 3 shows an overview of the GE language sites overlay and the pop-up window for a set of resources of the Yélî Dnye language resources provided by Levinson [20]. We expect that scholars from the physical and social geographical scientific community as well as those involved in development cooperation will be interested in this way of 'learning'.

5. Web portals

A different way of providing access to archived resources for non-specialist users are customized web-portals. These portals feature an attractive graphical design that is tailored to the speech community of a certain endangered language. Pre-defined search queries into the metadata catalog of the archive make it possible to create dynamic web content that will include archived resources if their metadata description matches the query. The queries can be hidden under buttons or links in the portals, so that users can execute them with a single click.

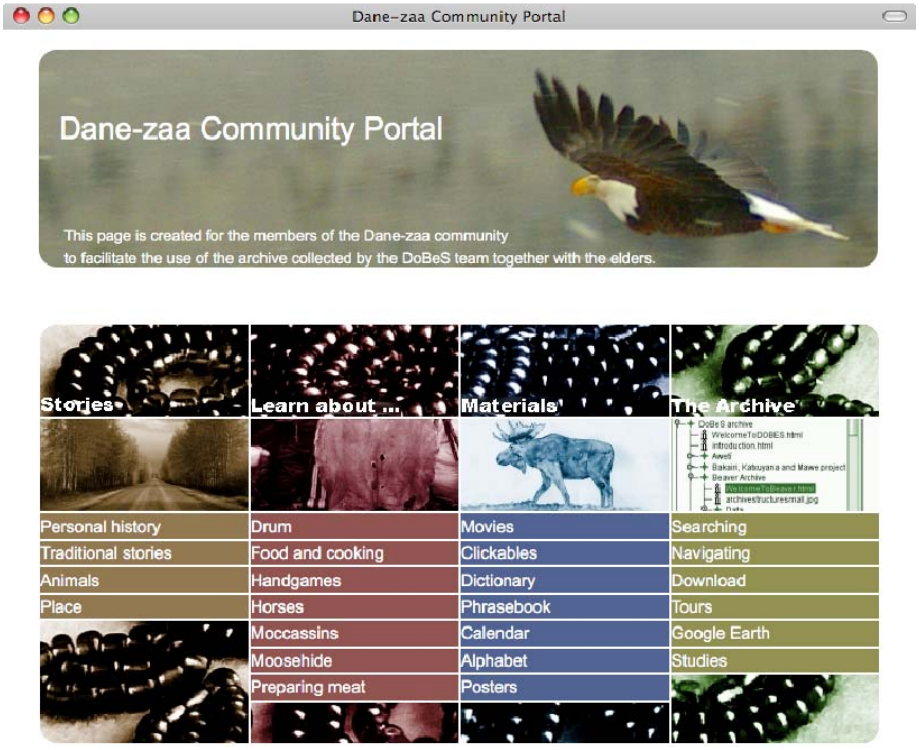

Figure 4: Dane-zaa community portal

Searching the IMDI metadata catalog from within a web portal is made possible by a Representational State Transfer (REST) Web Service [21]. The REST interface allows developers to specify rather advanced metadata queries and to get back results in XML format. On the portal side, the XML can be parsed and reformatted into lists in the same graphical style as the rest of the portal.

We have developed such portals for three of the documentation projects within DoBeS. An example is given in Figure 4 showing the community portal for the Beaver (Dane-zaa) language community in British Columbia and Alberta in Canada. It was designed by the DoBeS documentation team that is working on this language, and implemented by MPI staff using the Plone open source content management system [22] by adding a custom content type to execute and view the metadata queries.

6. Lexicon based access (LEXUS)

The MPI's Language Archiving Technology suite [23] comprises LEXUS, a tool for the creation and exploration of multimedia online lexica in the context of language description and documentation [24]. With LEXUS, researchers can easily enrich linguistic information with multimedia resources, for instance, to complement the abstract meaning(s) of a given lexical entry with a social or cultural context in which it can be used. Figure 5 shows an example of the Yélî Dnye lexicon [20]. In the left window frame, a word list gives access to the words in the lexicon; on the right side, the details of the selected lexical entry are displayed, including an image to illustrate one of the meanings of the word in question.

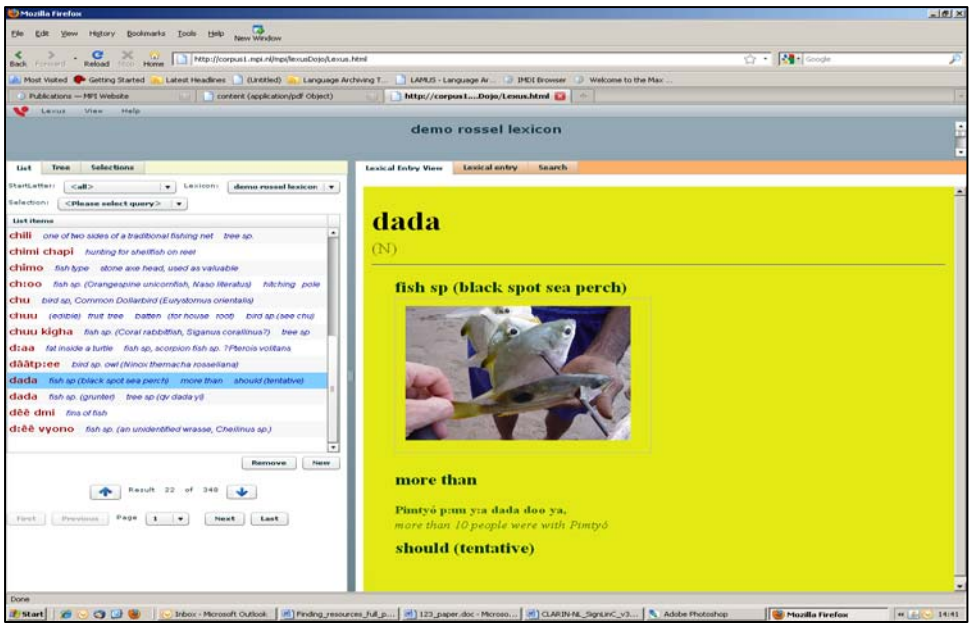

Figure 5: LEXUS Yélî Dnye lexicon [20], lexical entry 'dada'

Users can store multimedia elements 'locally' in the LEXUS database, or they can link – via a simple URL – to existing material in the MPI archive. In the latter case, LEXUS can thus be exploited as a structured navigational aid to archived material. The encoding of the URL determines whether LEXUS merely displays an image, or whether an

external viewer is opened to play (a segment of) an audio and video file together with existing, multi-tiered annotation (see Figure 6, displaying MPI's ANNEX tool [25]).

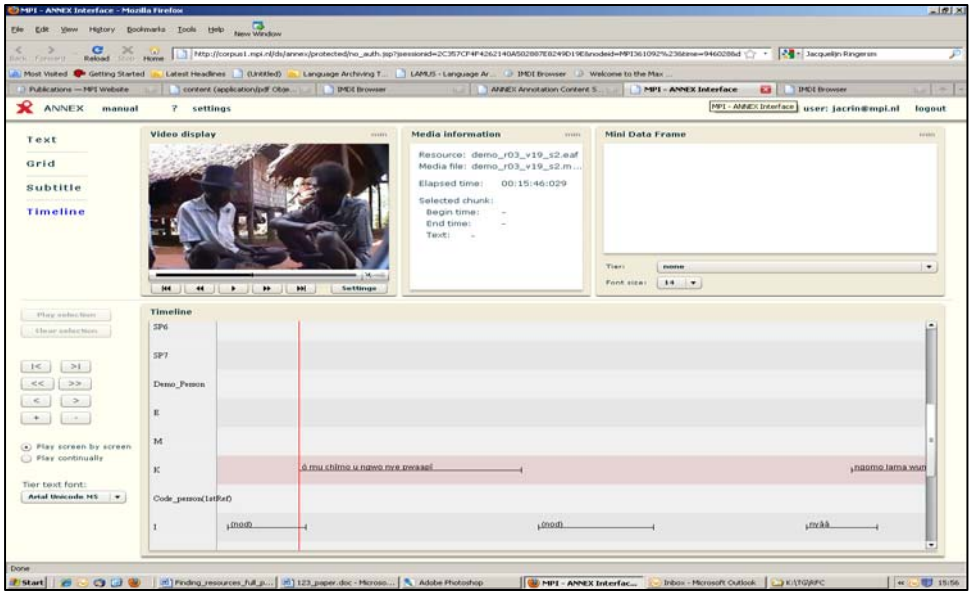

Figure 6: Annex player showing annotation and media file

In the DoBeS/Marquesan documentation project, multimedia linking is used to create image galleries. Here, exemplary frames from lengthy video files (which take quite some time to load in areas with slow internet connections) were selected and presented to show, for instance, a breadfruit preparation process. As the notion of breadfruit plays an important role in Marquesan daily life, religion and cultural activity, various lexical entries now link to the gallery (or some selected frame) to illustrate the rich vocabulary around this notion and preparation process. Galleries thus provide another method for grouping together and describing linguistic and extra-linguistic data.

7. **Ontology browsing (ViCoS)**

While expert users profit from LEXUS's access to archived material via linguistically-motivated categories, non-linguists require different guidance. Members of indigenous speech communities find access via linguistic means (LEXUS) or metadata descriptions (IMDI) of little help. They cannot easily find the richness of their language (and culture) in the mass of linguistic parlour and archive trees full of metadata. The ViCoS tool aims at addressing the need for a more conceptual access point to archived resources. ViCoS is designed to add value to LEXUS by allowing users to create arbitrary relations between lexical entries or their parts, such as images or example sentences [25]. The resulting network of semantic relations between entries form a kind of conceptual space, where words denote increasingly complex concepts. While each concept is anchored in the lexical space, its meaning becomes culturally relevant in the context of the concepts it is connected to.

ViCoS has been extensively used in the DoBeS project 'Iwaidja'. The Iwaidja lexicon, managed by LEXUS, consists of approximately 3500 lexical entries, mostly

nouns classified according to the linguistic information unit *semantic domain* to which they belong (animals, body parts, medicine *etc.*). With ViCoS, a conceptual space of shell fish and related objects has been created to complement the existing lexical space.

In Figure 7 we show the conceptual space for the Iwaidja concept 'rarlwa', which denotes the generic term for ‘oyster’. Three different types of relations are visible in the right upper corner of the window. The first one is 'eats' (in red line). In the conceptual space, the three blue colored concept nodes represent types of birds that eat 'rarlwa'. The second relation is of the 'is-a-kind-of' type (in black line). The group of eight purple colored concept nodes have 'rarlwa' as hypernym and the darker purple concept node ('jidingunda') is a hyponym of 'rarlwa'. Finally there is a 'sounds' relation (green dotted line), a non-semantic relation indicating that the phonological representation of the two concepts is similar.

Clicking on one of the nodes will place the selected concept in the centre, which is how users can browse the complete conceptual space. Each concept is anchored to (part of) a lexical entry (as maintained by LEXUS), and users can easily open LEXUS to show a concept node in its linguistic context. Moreover, ViCoS also allows users to create links between concepts of the conceptual space to other resources on the Internet.

The example shows the strength of the ViCoS tool: users can specify arbitrary relation types, objects can be grouped and colored, concepts are related to the word in the lexicon and ViCoS provides a direct link to the words as well as to the media resources in the archive, and finally ViCoS connects the object to other online resources.

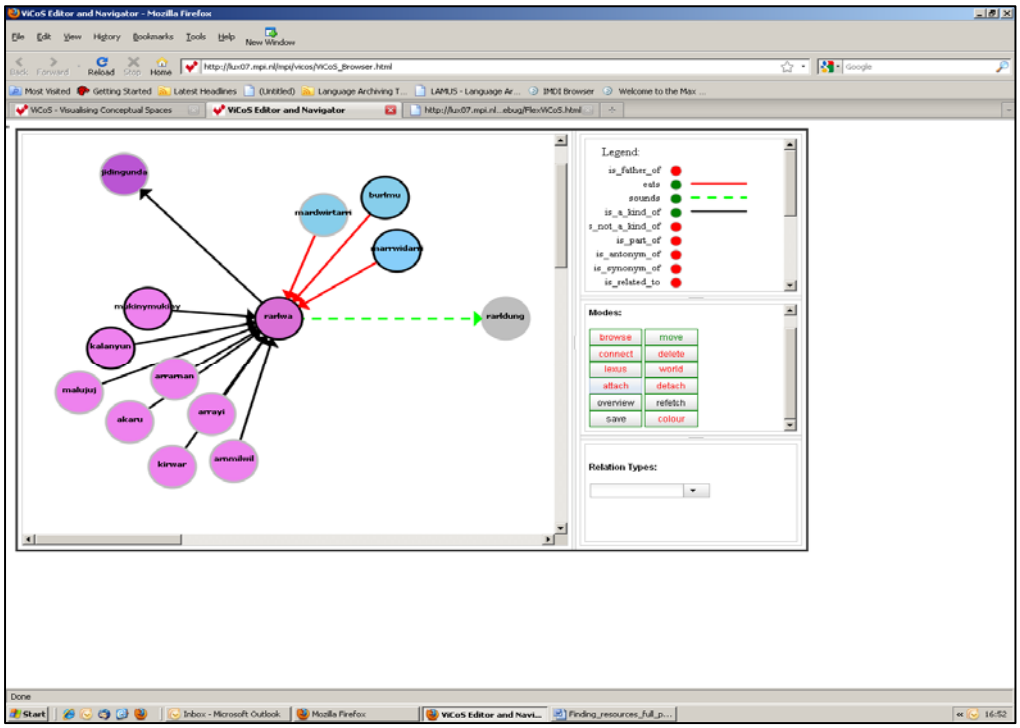

Figure 7: ViCoS browser, showing the conceptual space for the Iwaidja concept 'rarlwa'

Synthesis and conclusion

The MPI archive for linguistic resources hosts language data that originated from a diverse set of projects, each with different aims and perspectives ranging from language description and documentation to theoretical linguistic studies. Its data comprises research data from the institute's scientists, but notably also data collected in larger enterprises such as the DoBeS project. The archive is open to resources from other institutions, and a large amount of metadata is harvested on a regular basis to increase its visibility, persistent accessibility and to secure long-term data storage and preservation. Users of the archive are of very diverse backgrounds, ranging from Spinoza prize winning researchers to almost illiterate members of the speech communities.

When the archive was first initialized around ten years ago, the MPI provided only one method of accessing its data, namely, through the IMDI metadata browser and search engine. While the IMDI browser is still the tool of choice for MPI and external researchers, other user groups felt the need for tools that require less knowledge of the IMDI metadata set, and the archive's structure and content. Over the last three years we have started to explore and develop alternative ways to facilitate access to archive resources for target groups other than linguists.

The facetted browser that we have started to develop will provide access to all of the MPI archive via less than ten facets such as country, language, and organization, complemented with a full text search capability. The selection of three to five facets is usually sufficient to pinpoint to the resources users want to access. Given the increasing penetration of facetted browsing environment on popular websites such as *Europeana* and *Ebay*, we expect users experiencing little if any problems with this technology to browse the archived data.

We provide geographic browsing by which users access resources through a Google Earth overlay, which associates overlay placemarks with resources of a specific catalogue or language. We believe that geographic browsing is particularly attractive for members of the general public, given the overall success of Google Earth and the fact that non-expert users prefer to search for languages via language names or geographic locations.

Members of the speech communities of the documented (and endangered) languages get access to their data via purpose-built web portals. The design of the user interfaces varies across speech communities, and takes their particular needs or wishes into account. The portals' back-end makes use of IMDI metadata search services; here, pre-fabricated queries are initiated with specific user requests, sent to IMDI, and the result set is then presented to the user in their preferred thematic style and organization.

Finally, we provide access to archived data via lexical and ontological dimensions, using LEXUS and ViCoS, respectively. LEXUS allows users to browse parts of the archive via the words in the lexicon. As the views for word lists and lexical entries are configurable, LEXUS can be used by experienced linguists and novice users alike. With ViCoS, it is possible to complement lexical spaces with ontological ones. ViCoS offers indirect access to resources stored in the archive. And, as arbitrary URLs can be attached to a concept node, it is not only possible to link directly from conceptual

spaces to the MPI archive but to any resource that is accessible via a URL. ViCoS is being developed in close relation with researchers and speech community members of a DoBeS project. At a LEXUS/ViCoS training at a recent DoBeS meeting at the MPI in Nijmegen, ViCoS received very positive feedback from all attendants, and is already used in various pilot projects, each having constructed substantial conceptual spaces.

The mere storage and preservation of data in an archive should not be a goal by itself. At the MPI, we are attempting to make the resources easily available, accommodating the needs of the users and communities. The feedback from our users tells us that we are working in the right direction.

References

[1] Wittenburg, P., Skiba, R., & Trilsbeek, P. (2005) The language archive at the MPI: Contents, tools, and technologies. Language Archives Newsletter, 5, 7-9.

[2] Wittenburg, P. (2003) The DOBES model of language documentation. Language Documentation and Description, 1, 122-139 (also http://www.mpi.nl/dobes/)

[3] Crasborn, O., Zwitserlood, I. & Ros, J. (2008) Het Corpus NGT. Een digitaal *open access* corpus van flimpjes en annotaties van de Nederlandse Gebarentaal. Centre for Language Studies, Radboud Universiteit Nijmegen. URL: http://www.ru.nl/corpusngt/.

[4] Boumans, L.P.C & Crevels, E.I. (2005) Dutch bilingualism database. ACLA CLAN Workshop, Nijmegen (MPI) http://repository.ubn.ru.nl/handle/2066/41599

[5] ESF - Second Language Database http://www.esf.org

[6] Hoff, B.J. (1968) The Carib Language, Phonology, Morphology, Texts and Word Index. Verhandelingen van het Koninklijk Instituut voor Taal-, Land-, en Volkenkunde (Royal Institute of Linguistics and Anthropology) Vol. 55, Martinus Nijhoff: The Hague..

[7] Hippisley, A. (2000) Predicting the past: reconstructing the Slavonic colour lexicon http://corpus1.mpi.nl/ds/imdi_browser/?openpath=mpi318941%23

[8] UNESCO: http://www.unesco.org/culture/en/endangeredlanguages

[9] Declerck, T., Broeder, D., Romary, L., Uneson, M., Strömqvist, S., & Wittenburg, P. (2004) A Large Metadata Domain of Language Resources. In: Ferreira, F., Costa, R., Silva, R., Pereira, C., Carvalho, F., Lopes, M., Catarino, M. & Barros, S. (Eds.). *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004, Lisbon).

[10] Wittenburg, P., Broeder, D., Kemps-Snijders, M., Dimitriadis, A. & Soddemann, Th. (2007) A Federation of Language Archives Enabling Future eHumanities Scenarios. *German E-Science Conference* (GES 2007, Baden-Baden).

[11] Access procedures: http://www.mpi.nl/DOBES/archive_access/access_procedures

[12] Wittenburg, P. (2004). The IMDI metadata concept. In S. F. Ferreira (Ed.), Workingmaterial on Building the LR&E Roadmap: Joint COCOSDA and ICCWLRE Meeting, (LREC2004). Paris: ELRA - European Language Resources Association.

[13] Flamenco Facetted Browser: http://flamenco.berkeley.edu.

[14] Apache Solr: http://lucene.apache.org/solr.

[15] Google Earth: http://earth.google.com/

[16] Van Uytvanck, D., Dukers, A., Ringersma, J., & Trilsbeek, P. (2008). Language-sites: Accessing and presenting language resources via geographic information systems. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).
[17] Google Earth Overlay: http://www.mpi.nl/DOBES/dobesmap/language_sites.kmz
[18] Paradisec: http://www.paradisec.org.au/
[19] AILLA: http://www.ailla.utexas.org/
[20] Levinson, S. (2006). Enrolling other sciences in language documentation: describing an isolate language in Papua New Guinea. DGFS Conference, Bielefeld.
[21] Fielding, R. T. & Taylor, R. N. (2002), "Principled Design of the Modern Web Architecture" (PDF), ACM Transactions on Internet Technology (TOIT) (New York: Association for Computing Machinery) 2 (2): 115–150, doi:10.1145/514183.514185
[22] Plone: http://plone.org
[23] LAT: http://www.lat-mpi.eu
[24] Ringersma, J., & Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme & R. van Son (Eds.), Proceedings of Interspeech 2007 (pp. 1529-1532). Adelaide: Causal Productions.
[25] Berck, P. & Russel, A. (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)*
[25] Zinn, C. (2008). Conceptual spaces in ViCoS. In S. Bechhofer et al. (Eds.), The semantic web: Research and applications (pp. 890-894). Berlin: Springer.