# Ontology-based Language Archive Utilization

**Article**

**5 authors**, including:

Peter Berck
Radboud University
**19** PUBLICATIONS **363** CITATIONS

SEE PROFILE

Hans-Jörg Bibiko
Max Planck Institute for the Science of Human History
**9** PUBLICATIONS **435** CITATIONS

SEE PROFILE

Peter Wittenburg
Max Planck Institute for Psycholinguistics
**91** PUBLICATIONS **796** CITATIONS

SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**

Project    The Language Archive View project

Project    Implicit Linguistics View project

# Ontology-based Language Archive Utilization

**Peter Berck, Hans-Jörg Bibiko, Marc Kemps-Snijders, Albert Russel, Peter Wittenburg**

MPI for Psycholinguistics, MPI for evolutionary Anthropology
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
{peter.berck, marc.kemps-snijders, peter.wittenburg}@mpi.nl, bibiko@eva.mpg.de

### Abstract

At the MPI for Psycholinguistics a large archive with language resources has been created with contributions from many different individual researchers and research projects. All of these resources, in particular annotated media streams and multimedia lexica, are accessible via the web and can be utilized with the help of web-based utilization frameworks. Therefore, the archive lends itself to motivate users to operate across the boundaries of single corpora and to support cross-language work. This, however, can only be done when the problems of interoperability, in particular at the level of linguistic encoding, can be solved in an efficient way. Two Max-Planck-Institutes are cooperating to build a framework that allows users to easily create their own practical ontologies and if wanted to relate their concepts to central ontologies.

## 1. Introduction

In addition to the well-known language resource providers LDC (Philadelphia) and ELDA (Paris) we see the emergence of an increasing number of language resource archives such as MPI (Nijmegen), BAS (Munich), TST Center (Leiden/Gent), SOAS (London), AILLA (Austin), Paradisec (Sydney) during the last years that cover highly valuable material and that turn over to offer web-based utilization interfaces. For all these archives it is true that they store different linguistic resource types such as annotated media files, lexica, sketch grammars etc and that this material was created and deposited by different individuals or projects. This implies that the level of heterogeneity at the technical and linguistic encoding level is fairly high. In this paper we will not focus on the aspects of unification at the technical encoding level (character encoding, structure/format issues), since these aspects are tackled by a move to more generic standards such as UNICODE [1], LAF [2] and LMF [3].

We will focus on the design of a framework that will allow researchers to easily bridge the gap created by the differences in encoding linguistic phenomena. Currently, it is effectively not possible for a normal linguist to carry out searches for example that include contributions from different projects or to easily link a lexicon with an annotation. Although metadata is part of the linguistic encoding problem, we will also not discuss this aspect in this paper. It has already been described that metadata is about limited and stable vocabularies so that mappings between OLAC [4] and IMDI [5] for example could easily be created. With respect to the encoding of the content – be it as tiers in structured annotations or as attributes in lexica together with the values these can take – we are faced with a extensively large and dynamic vocabulary. Their usage is dependent on the languages being studied, on linguistic theories and on the purpose of the research in mind. Since linguistics is still a developing field of research this dynamic character of linguistic encoding will not change.

## 2. Central Ontologies

Of course, we expect that top-down generated ontologies or knowledge components such as GOLD as it was created within the E-Meld project [6] or the Data Category Registry as it is being developed in ISO TC37/SC4 [7] will improve interoperability in the long run. But this will depend on a number of factors indicating that it will take a while until we can take real profits: (1) Tools have to be available and used that support the interaction with such components to allow an easy re-usage of concepts. (2) The components have to be established and finalized and have still to demonstrate whether linguists will be willing to use the included concepts. In general, linguists want to stick with their concepts that to a certain extent include very detailed definitions and reflect their theoretical convictions. (3) For the huge amount of language resources that already exist and that are created at this moment there is no direct link with concepts included in these ontologies. Some projects comply to the TEI [8] or EAGLES [9] standards/ recommendations, but these are the exceptions.
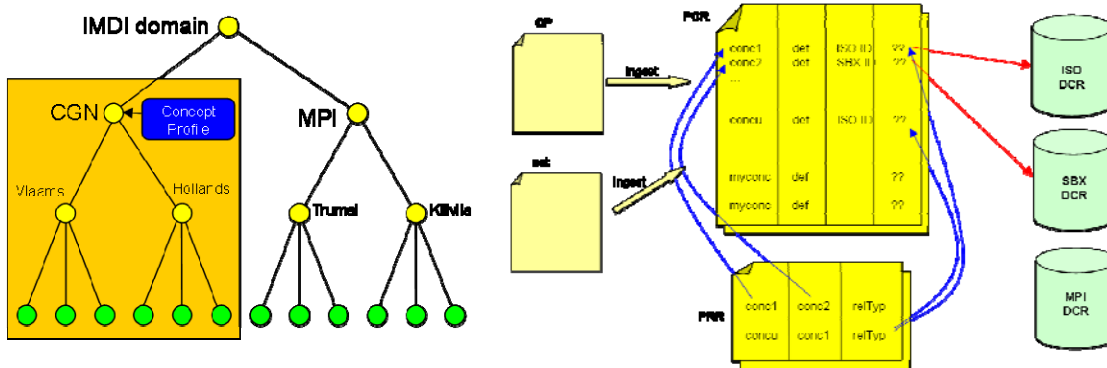
Metadata schemas that relate the vocabularies used in a concrete language resource with central ontologies could be used, but their creation costs much time creators will in general not willing to invest and, above all, this only makes sense when many researchers will do the same. If mapping schemas are only created incidentally then this will not lead to an interoperable domain.

## 3. Bottom-up Driven Ontologies

In contrast to this approach we will rely on a bottom-up strategy, i.e., we assume that researchers who want to carry out some deep analysis on some resources will necessarily invest the time to create the necessary mapping files. Where possible we will create concept profiles which are registries of concepts used in a sub-corpus. These concept profiles are used by following the recommendations of ISO for the creation of entries in the data category registry. Of course, we do not expect the individual researcher to fill in all language specific fields provided by the ISO DCR framework. But we should take care that this kind of semantic information can easily be exchanged. So, for example, for the Dutch Spoken Corpus

project [10] all concepts used were defined carefully. Also for the teams that do language documentation work [11], for example, there is a duty to describe the concepts they are using.

We suggest to provide a framework that makes it easy to integrate such existing concept profiles where already existing, to create concept profiles containing the necessary concepts where necessary, to create mappings between the various registered concepts and to publish all components (concept registries and mappings) so that they can be re-used by others. The core is an editor that easily allows to gather, combine, store and share concepts that are defined by the individual researchers and registered primarily for their own purposes. Of course such an editor should allow the user to make use of existing mappings to central ontologies or create new ones. The following figures give an indication of what is currently being developed.

specified precisely. From previous work in ECHO [13] and many discussions in the field it does not make sense to provide more detailed operators such as they are defined in OWL [14], for example. We expect that most of the users who want to simply define a search space will not take the time to deeply elaborate on the semantic relationship between two concepts, but rather use the "mapsTo" type to exploit the content for their purposes. A number that may indicate the degree of fit between two concepts may be used to calculate a rating that may be used to rank the results. But it is too early to make statements about the feasibility of such ideas.

## 5. Utilization Frameworks

The work at the MPI for Psycholinguistics is driven



*The left figure shows how concept profiles can be associated with corpora in an IMDI organized archive. At the root node of the corpus such a concept profile can be stored such that a crawler would always find it when a resource from that corpus is going to be used. In the right figure it is indicated how concept profiles, sets of concepts created on the spot can be integrated to a Personal Concept Registry and how mapping can be carried out by making use of a Personal Relation Registry and mappings to central data category registries. Of course, such PCR and PRR can be uploaded into the archive at certain places to allow other users to re-use them.*

A complete architecture was designed by the MPIs for Psycholinguistics and evolutionary Anthropology. A first editor version was developed and will now be integrated and tested. For all concept registries a subset of the ISO DCR structure will be used opening the possibility for an easy exchange of information. For relations at first instance a simple XML structure will be used with the option to also create an RDF instantiation [12].

## 4. Relation Types

With respect to the creation of mappings we have to anticipate the usage scenarios. Mappings will be created not only be based on underlying linguistic theories, but in particular based on concrete and pragmatic wishes of the researcher. Therefore, it seems to be wise to limit ourselves at the beginning to a few simple relation types such as "isEquivalent", "isSubclass", "isSuperclass" and "mapsTo". While the first three are logically exactly defined, the latter is an indication of a semantic overlap that cannot be

by two web-based utilization frameworks that are now available (ANNEX [15], LEXUS [16]). Both allow to carry out for example searches on various resources selected from the large archive. Currently, they allow the user to select attributes or tiers and associate patterns with them. This is the first most simple version of creating interoperability. They can be stored to be able to re-use them, but they are too inflexible. The following figure indicates the kind of functionality offered at this very moment in ANNEX and LEXUS.

## 6. Conclusions

At the MPI web-based technology was developed to allow users to utilize the archival content. Users can select arbitrary resources via metadata searching or browsing and carry out searches on them or to compare them etc. At the level of character encoding we assume UTF-8 encoding and at the level of structure we support "generic" formats for lexica (LMF) and for annotations (EAF). All lexical and annotation structures we know of

can be converted to these two formats, however, we also support Shoebox, CHAT and Transcriber formats.

another dimension of achieving interoperability. We expect that only a few users will use these features at the
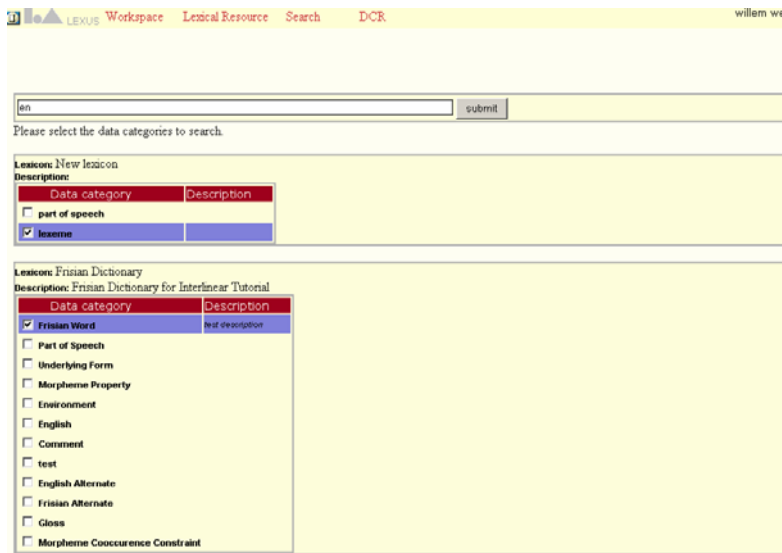


*This figure shows a screenshot from the search component of LEXUS tool indicating that the user can search across different lexica (here: New Lexicon and Frisian Dictionary). However, at this moment the user has to select the attributes from the different lexica for what a search pattern has to be specified. A step towards using bottom-up defined ontologies is necessary to facilitate cross-lexicon searching.*
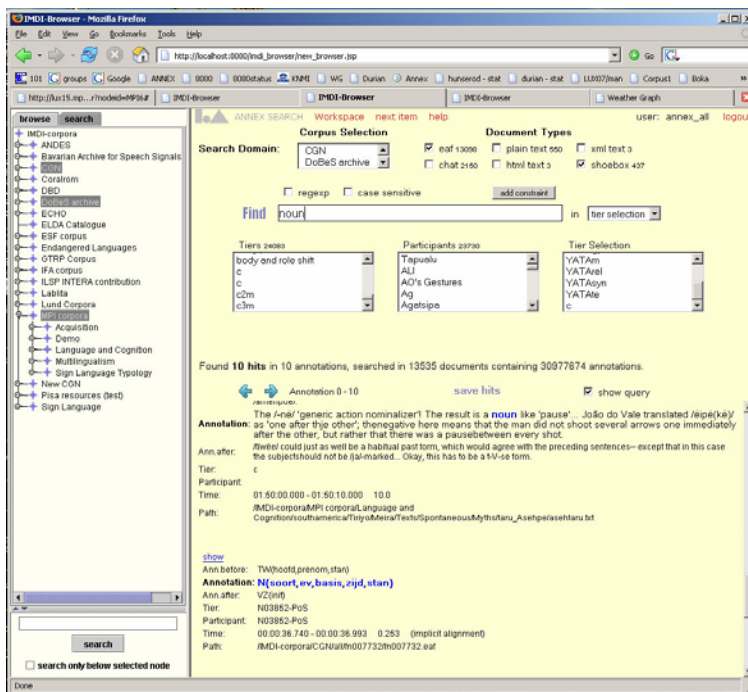


*This figure shows a typical screenshot when carrying out cross-corpora searches with ANNEX. The user selects a number of corpora or resources within corpora, defines a search patterns and carries out a search that includes all selected resources. For the case of simplicity we show a search for "noun" on a number of selected tiers. The user could add another constraint and search on other tiers for example for "no". Implicitly it is stated therefore that "noun" and "no" should be treated the same for this particular search. A step towards using bottom-up defined ontologies is necessary to facilitate cross-lexicon searching.*

However, at the level of linguistic encoding the differences in terminology have to overcome. Currently, the users will be offered tiers and attributes which they can select and for each of them they can specify a pattern. This can only be seen as a first step. The next step has to be a framework that allows a user to easily select concepts (at the tier/attribute and at the value level), to easily specify relations between them and mechanisms to make these "personal practical ontologies" persistent, sharable and re-usable. The two MPIs created a complete design and are currently implementing this bottom-up driven concept.

Finally, users will have the possibility to also draw relations with central ontologies and, therefore, open

beginning, since creating sharable and re-usable knowledge components will cost them quite some efforts. In summer 2006 we expect that a full-fledged version is operational and then we have to see how this will be used. Already now we are convinced that only a combination of bottom-up and top-down created ontologies will help the field.

## 7. References

[1] http://www.unicode.org/
[2] http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf
[3] http://www.lrec-conf.org/lrec2006/IMG/pdf/OutlineForLMFTutorial.pdf

[4] http://www.language-archives.org/
[5] http://www.mpi.nl/IMDI
[6] http://emeld.org/gold-ns/index.cfm
[7] http://www.tc37sc4.org/new_doc/ISO_TC_37-
    4_N133_DCR_for_TC37.pdf
[8] http://www.tei-c.org/
[9] http://www.ilc.cnr.it/EAGLES/
[10] http://www.tst.inl.nl/producten/CGN/
[11] www.mpi.nl/DOBES
[12] http://www.w3.org/RDF/
[13] http://www.mpi.nl/ECHO
[14] http://www.w3.org/TR/owl-features/
[15] http://www.mpi.nl/annex
[16] http://www.mpi.nl/lexus