On-line Cooperation for Language Documentation, Research and Development with XNLRDF: The Example of Formosan Languages

Oliver Streiter
National University
of Kaohsiung

Simon Chun-feng Su
National Sun Yat-Sen
University

Leonhard Voltmer
European Academy
Bolzano/Bozen

Yoann Goudin
National Cheng Kung
University

Peripheral languages need for their promotion, their documentation, for linguistic research and for the development of NLP and e-learning resources a showcase, a counter, a well organized depot and a supply chain. The XNLRDF project attempts to provide all this in an Internet-based platform, trying to overcome the respective deficiencies of related projects. The main difference between XNLRDF and apparently similar projects is that the data in XNLRDF are structured, computational data. In addition, these data can be created, modified, compiled and downloaded freely through a web-based interface, allowing professional and non-professional experts to collaborate. The impact of this approach will be discussed, using the examples of Formosan Languages.

1 Introduction

Peripheral languages in general and endangered languages more specifically need for their promotion, their documentation, for linguistic research and for the development of NLP and e-learning resources a rich infrastructure. This infrastructure has to fulfill a number of tasks, such as the transfer of approaches, of data structures or tools from one language to another, so that what has been designed for one language can be of use for another. In addition, the infrastructure should allow linguists and native speakers to collaborate (cf. Eisenlohr 2004), to draw public attention to the language in question, to enhance awareness, identity and pride among speakers, to collect, enrich and archive language data and, ultimately to provide useful tools and data which support the interrelated activities of linguistic research, language documentation, language promotion and language learning.

The XNLRDF project attempts to set up an infrastructure for this through an Internet-based platform, trying to find ways how to compensate for deficiencies of related projects. Such projects are (a) world language databases, the best known of which is the Ethnologue database, (b) language documentation projects and language archives that provide coarse grained information on a limited number of languages and (c) software pools which provide electronic resources such as dictionaries or spell checkers for otherwise unrelated languages. Without questioning the importance and achievements of all these projects, we will start in Section 2 with an analysis of what we feel are their respective shortcomings. From these shortcomings we will derive in Section 3 the ideas which shaped the basic design of XNLRDF: 3.1 Structured data, 3.2 Collaborative data creation, 3.3 Computational data, 3.4 Free data and 3.5 Tools and utilities. In Section 4 we will than describe the basic data structure of XNLRDF. Section 5 will focus on the impact XNLRDF might have on Austronesian Languages in general and Formosan languages in particular. Section 6 will be dedicated to the discussion of problems we encountered with this approach. A summary and discussion will round up this contribution.

2 World Language Databases, Language Archives and Software pools

In this section we shall describe three types of language resources, which share the basic common purpose, i.e. making language data available. These are world language databases, language archives and software pools. From a critical appraisal of these systems, we will proceed to the formulation of the principles which shaped XNLRDF, a database under permanent development. The design of this database tries to overcome the limitations of other approaches to language data preservation and proliferation.

2.1 World Language Databases

The term ‘world language database’ is not a standard term. We will use it here as a cover-term for systematic language descriptions which contrary to language archives, linguistic publications or other language-related resources do not focus on particular languages but potentially describe all languages of the world. Providing information for hundreds or thousands of languages, such databases trade necessarily a broad coverage for a shallow language description. Some of them concentrate on one dimension of linguistic description only, e.g. morpho-syntactic properties, writing systems etc. Most of the world language databases can be consulted on-line and we will limit our analysis to only these.

Language databases usually use one or several identifiers for a language and relate to it data, descriptions, documents or downloads. Some world language databases provide information about the number of speakers, the language family, the writing system, typological features, language variants or where and when spoken. Some others provide texts, bible translations, word-lists, audio files, linguistic articles or references to teaching material or Web-pages. If no language archive has been compiled for a specific language, which may be the case of more than 95% of the languages of the world, world language databases are the first and sometimes the only source of information that can be consulted besides the fragmented information that one might find in libraries or on the Web. Due to the shallow description in world language databases, complete information about a language, more often than not, is unfortunately scattered over different world language databases, with overlapping or contradictory information in each. As these databases do not necessarily share a common conceptual framework nor a common set of meta-data, they are accessible only by human experts who have to guess their interrelation.

Unless experts extract data from traditional linguistic publications and merge them in world language databases, the data are lost when the supporting material decays or becomes inaccessible (cf. Streiter, Scannel and Stuflesser 2006). Unless the description of different world language databases are merged in larger ones, their interrelatedness remains implicit and the data cannot be used in research that requires the resources of more than one world language database.

Combined world language databases actually would allow to compare the writing systems of the world, to analyze when and why people shift from one language to another or to analyze the relation of typological features and their local distributions (e.g. Haspelmath et al. 2005). Facilitated through easy-to-use spreadsheets, GIS (Geographic Information Systems) (c.f. Streiter 2007) and powerful statistical programs such as R (http://www.r-project.org, retrieved 2007-05-20), such research questions became feasible for a wider community and language questions can be handled at a lower level, including the people directly concerned with such questions. We therefor predict that world language databases will play an ever-growing role in linguistic research, especially if a question cannot be answered by analyzing one language. However, how can world language databases be developed, maintained and validated for 8000 languages and their variants, their 100.000 writing systems and their complex historical, geographical and cultural relations? Can we identify general trends that have emerged in the history of these databases in order to cope with these challenges?

2.1.1 The Ethnologue Database **(http://www.ethnologue.com,retrieved 2007-05-16)**

This project, which started in the 1950s, is currently the most complete and most frequently cited world language database. Most other language databases derive their meta-data (the language names, language codes) from the Ethnologue database. Identifiers however changed recently so that some derived databases use the Ethnologue 14 codes, while others use the codes of Ethnologue 15.

As it is the oldest one, it is based on methods of data management that are no longer considered up-to-data. The Ethnologue database provides semi-structured data on the world's languages, language maps, bibliographical data and computer resources like fonts and programs for fieldwork in flash-card-like entries. The fields in these flash-cards however are badly defined. The population field, for example, might store concepts as different as the size of the ethnic population associated with a language or the number of native speakers of a language. The staccato-like prose within the fields creates scoping problems, e.g. when dropping the subject of a sentence. Examples below will illustrate this. Terms in this database are, in addition, ambiguous and the same term might refer to a region, people of that region, a language of that tribe, a dialect of that language and the language group of that language. The example of the entry for the language Argobba at http://www.ethnologue.com/show_language.asp?code=agj illustrates some of these problems:

Population 10,860 (1998 census). 44,737 monolinguals. Population includes 47,285 in Amharic, 3,771 in Oromo, 541 in Tigrigna (1998 census). Ethnic population: 62,831 (1998 census).

Region Fragmented areas along the Rift Valley in settlements like Yimlawo, Gusa, Shonke, Berket, Keramba, Mellajillo, Metehara, Shewa Robit, and surrounding rural villages.

Dialects Ankober, Shonke. It is reported that the 'purest' Argobba is spoken in Shonke and T'olaha. Lexical similarity 75% to 85% with Amharic.

Classification Afro-Asiatic, Semitic, South, Ethiopian, South, Transversal, Amharic-Argobba

Language use 3,236 second-language speakers. The ethnic group near Ankober mainly speaks Amharic; the group near Harar mainly speaks Oromo. The ethnic group is working to foster ethnic recognition. Speakers also use Amharic or Oromo.

In this entry, if the ethnic population (which one?) is 62.831 and the number of monolinguals in Argobba is 44.737 what does 10.860 refer to? And what do the references to languages 'in Amharic', 'in Oromo' etc mean? 'Oromo' in addition can refer to the language or the people. Human reasoning is required to convert these data into something meaningful. With the data on the Region we observe the general lack of standard identifiers, temporal or geo-spatial references. Scoping problems in the statements are omnipresent. What do the rural villages surround? The Rift Valley or the settlements? Which dialect does the lexical similarity refer to? Ankober, Shonke, the purest Argobba or Argobba? What does '3.236 second-language speakers' mean? 3.236 of the ethnic population speaking a second language, or 3.236 of another mother tongue speaking Argobba as second language?

The legal status of the data is uncertain, as these data have been drawn from thousands of linguistic publications. Thus, while the publications in print version or on the web might be copyrighted, the individual pieces of information contained in them cannot be copyrighted, i.e. there is no copyright on SVO of English. The web-version of the database is updated every 4-5 years, which, compared to Wikipedia is extremely slow. In addition, these updates create flat, non-historic language profiles, e.g. when replacing the number of speakers with newer data. A researcher who wants to use these data for tracing the development of languages in numbers of speakers, thus would have to go back to older versions, which however are no longer available on the web and which in addition have different identifiers. This view without historical foundations leads furthermore to inconsistencies. Languages are located in Russia with census data from the Soviet period. If not contradictory, either the term Russia is ambiguous and refers to the Russian Federation or the Russian SFSR or Imperial Russia, or it is even ambiguous and refers either to a political unit or a geographical area.

The database is basically an English-language database and language names others than in English or autonyms are not provided, even not if the most relevant scientific language for the description of a language in an entry is like Mandarin, Spanish, Russia or Indonesian a widely spoken scientific language. Characters others than Latin characters are systematically avoided.

To sum up, it thus seems that the wealth of the Ethnologue database should be subjected to a systematic transformation to fulfill the requirements of modern terminology databases. This however cannot be achieved automatically and would require human intervention in most cases. This is what XNLRDF, among others, aims at.

2.1.2 Omniglot (http://www.omniglot.com, retrieved 2007-05-15)

The only marginal references to writing systems in the Ethnologue database is partially compensated for by the Omniglot database. This database represents the heroic work of a single person, Simon Ager, collecting detailed information about the writing systems of more than 400 living and ancient languages. The prose texts are generally more wordy and less ambiguous. The automatically harvesting of information from these resources however therefore is more difficult. In addition, the data provided, such as textual examples or characters are frequently given as images, limiting the usefulness of this as a resource for computational purposes, e.g. for extracting a list of characters for a given writing system. We will discuss this aspect below as *computability*.

The data thus seem to have been created like an encyclopedia to be consulted by humans. Data can be accessed through an A-Z index, (writing) direction index and language index and a classification of writing systems (e.g. alphabets, abjads, abugidas, etc).Though it does not provide fonts or other computational resources, direct access to such resources is frequently given through links. The main function, of this database is thus to provide a descriptive prose text about languages and their writing systems on the one hand, and a kind of structured access to data outside the database on the other. The information in Omniglot are generally considered to be trustworthy, however neither Ami, Bunun, Tayal, Tsou or Paiwan can be found in this database, showing the work of single persons, even that of a hero, is necessarily limited.

2.1.2 Wikipedia (http://{en,de,jp,fr,zh,ru,..}.wikipedia.com, retrieved 2007-05-16)

Wikipedia contains descriptions of languages, writing systems and ethnic groups associated to languages. As the articles in Wikipedia are written by thousands of experts in collaboration, Wikipedia in a short time overtook Omniglot in coverage and depth. In most cases, languages are described more in detail than in the Ethnologue database, but for many marginal languages, we basically find only a copy of the Ethnologue data. Language data, such a characters, phonemes, example sentences etc are given in Unicode and not, as in Omniglot, as images, making these data available for computational purposes.

Thus, through its collaborative nature, Wikipedia easily seems to overtake other, older language databases, while still sticking to one of their, according to our needs, main disadvantages, i.e. providing information in mainly unstructured formats. Exceptions are loosely formatted translation lists, phoneme-lists etc. Entries in Wikipedia try to mention all available standards and thus makes data *interoperational* (language names, scripts, localities). In addition, the license under which the data in Wikipedia are published potentially allows for the usage of its data in other projects. This is one motivation for why people work voluntarily for Wikipedia. In addition, it facilitates the data perseverance as many people may copy the data and redistribute them.

Many times, different and partially complementary information can be found in the English, Mandarin, Russian, Taiwanese, German, French, Spanish etc. versions. This imposes advanced readings skill for a single user who wants to get a round picture of a language. In addition, from times to times, the information in parallel different language versions of Wikipedia is not complementary but contradictory or simply not coherent. The religion of a tribe might be described in one language as 'belief in spirits' (http://de.wikipedia.org/wiki/Blang, retrieved 2007-05-21), in another as 'animism' (http://en.wikipedia.org/wiki/Blang, retrieved 2007-05-21), in still another as 'polytheism' (http://es.wikipedia.org/wiki/Blang, retrieved 2007-05-21) or 'ancestor worship' (http://zh.wikipedia.org/wiki/%E5%B8%83%E6%9C%97%E6%97%8F, retrieved 2007-05-21). This is the price Wikipedia has to pay for providing unstructured data in more than one language. If structured data would be provided in many languages, this could be done by localizing a restricted vocabulary, keeping the data at the same time coherent and consistent. The respective advantages and disadvantages of a restricted vocabulary will be discussed below in more detail.

Again, this project is of enormous importance. Porting the data into a structured format, however, requires human intervention and a general structural framework in which these data can be stored. This is what XNRDF aims at. Although Formosan languages are covered in Wikipedia, the entries lack far behind the available knowledge on these languages, compare, for example, the Mandarin Wikipedia entry on Tsou (http://zh.wikipedia.org/wiki/%E9%84%92%E8%AA%9E, retrieved 2007-05-21) and Tung (1964).

2.1.3 Language Typology (http://www.unicaen.fr/typo_langues/consultation_langue.php, retrieved 2007-05-15)

The Project started 2005. This database tries to assign typological features to the meta-data taken from Ethnologue 14. Currently about 250 languages have been described partially, more information can be suggested by users through a Web-interface. User-entered data are then validated or rejected by internal staff members. Through this intermediate step, the advantages of having a large collaborating community on the one side, and the control of the coherence and consistency of the data are combined. The disadvantages however are a slowdown in information flow and a certain frustration with users how find their data wrongly rejected. The first author, whose mother-tongue is German and who as worked for years on formal models of German grammar, 's attempt to assign SOV to High German was rejected by the French-speaking staff members. As data on German verb position are widely available, the most likely reason for this might be that the stuff members apply different criteria for the assignment of such categories than the first author did. With other, marginal languages, we wonder how the stuff members will control the data, without doing a complex research themselves and thus having no advantage at all from the collaborative interface, thus how are they going to check the word order in Tsou or Bunun? In effect, for Ami, Bunun, Paiwan, Tayal and Tsou we find beside the data copied from Ethnologue no additional typological data.

This database thus suffers from the general problem, that data fields are not sufficiently defined. In principle, definitions of the data fields as we find them Haspelmath et al. (2005) would be required to define a category. This definition includes positive descriptions and negative descriptions (what is not counted as belonging to the phenomenon), the definition of possible values and, of course examples. This, admittedly, is barely possible in an on-line database directed at, among others, native speakers. In XNLRDF, we therefor link whenever possible the XNLRDF data categories to those of the GOLD 0.3 ontology (Farrar & Langedoen 2003, http://www.linguistics-ontology.org/gold.html, retrieved 2007-05-21). In addition, examples of filled-in categories in other languages or other writing systems are provided in XNLRDF through a click on the field name.

In addition, it seems that the construction of such a database and its interface could profit from insights in the construction of questionnaires such as has been acquired in social sciences (c.f. Mummendey 1995), e.g. asking already known things, asking the same thing in two different ways or offering impossible answers would allow to induce the reliability or motivation of a user that enters data. It is this strategy that we will pursue in the future for XNLRDF. It might be further possible to store with a category not only one value, but the distribution of values provided by different users.

The strong aspects of the Language Typology project are that data are kept in a relational database and can be downloaded in comma-separated value files. In addition, on-line programs can be run to test typological hypotheses. To sum up, although limited in scope this database already contains a number of features that we consider trend setting: The use of structured data in a relational database, the potential contribution of on-line users and the free download of the data. The attempt to control the data by local experts however does not seem to be promising. The situation might be different with a greater number of experts, each working on different language subfamilies.

2.1.4 ODIN (http://www.csufresno.edu/odin, retrieved 2007-05-15)

ODIN, the *On-line Database of Interlinear Text*, provides currently more than 40000 instances of linguistic data for 731 languages. The data are mostly short example phrases in the form of interlinear glossed text (IGT), drawn from linguistic articles. These IGTs are indexed by language name or Ethnologue code. The language instances can be searched by grammatical markup (e.g., NOM, ACC, ERG, PST, 3SG), and by linguistic constructions (e.g. passives, conditionals, possessives, raising constructions, etc.). The language data are harvested automatically from on-line sources and classified (semi-)automatically. The challenge is that IGTs in linguistic articles are written for humans and neither explicit in structure nor relation. The data has to be disambiguated through statistical taggers and parsers. The ODIN project, according to the creator,

> *... can be seen as a prototype for the linguistic search tools of the future, providing the facility to search across thousands of instances of language data in hundreds of languages, unified into a common format and normalized to a common vocabulary.*
>
> Lewis (2006:1)

This project is important in many respects. Scattered data available on the web are harvested and united into a common framework. Language identification, however, is done, unfortunately by the outdated Ethnologue 14 language code and does not descend to the level of dialects. Second, the grammatical markup is mapped onto the GOLD ontology (Farrar & Langedoen 2003), thus providing a uniform interpretation for the grammatical markup, even if the definitions in GOLD are only a fraction of that we find in Haspelmath et al. (2005). Also important is the combination of automatic techniques like Web-harvesting, tagging, aligning and the manual verification of the data. Unfortunately, mainly for legal reasons, the IGTs extracted from the articles are not available as computational data. Thus, although they theoretically could be transformed into an XML-format such as proposed by Simons et al. (2004), they are only provided as links back to the PDF-files containing them as pictures. Once the PDF-files are removed or displaced by the authors or the authors' institutions, the data will be lost. Although we doubt that any linguist would object his or her data to be reproduced , if appropriately cited, as life data, the strict US copyright law seems to forbid this. It goes without saying that any future project that could overcome the latter limitation would become a resource of enormous importance. The number of hits in ODIN for Formosan languages are Ami:0, Bunun: 0, Paiwan: 0,Tsou: 12, retrieved 2007-05-21.

2.1.4 The Rosetta Project (http://www.rosettaproject.org, retrieved 2007-05-21)

The Rosetta Project is a language archive which builds mainly on the Ethnologue data. As the Ethnologue database, the Rosetta Project focuses on languages and largely ignores writing systems. In addition to the Ethnologue backbone, a large community uploaded documents related to the languages in the form of image files. Other resources like audio and video-files, computational texts and annotations could be stored in this archive, but such resources don't seem to have been uploaded in great quantity. The image files are mainly scanned linguistic text books. However, only the first pages are scanned for copyright reasons. In addition, this kind of resource might in the near future be supplanted by the Google Book Search Library Project (http://books.google.com, retrieved 2007-05-21), with thousands of books scanned and indexed after character recognition. Book retailers like Amazon allow also to glimpse into books. As such, these image files do not create a new or distinctive type of resource. For Ami we found the translation of the Genesis, for Tayal one glossed text, for Paiwan 7 different resources and for Bunun , Tsou 0 resources (retrieved 2007-05-21), although with Google Book Search we find easily material (e.g. http://books.google.com/books?id=aC8sAAAAMAAJ&dq=Tsou&q=wash&pgis=1#search, retrieved 2007-05-21).

Thus, limited by copyright issues and having as main data resource only image-files that will be available in the near future to an even larger extent elsewhere, the Rosetta Project is currently developing into various directions, adding community tools and enhancing its database infrastructure by adopting RDF. The data-scheme however, as represented in Good and Hendryx-Parker (2006), is far to simplistic to replace the data in the unstructured images in the near future by structured data. There is a huge gap between the data available as image and the data-structure designed and implemented. XNLRDF tries to improve also on this, giving a much richer and steadily growing data structure in which a much larger part of the available unstructured data can be described formally.

2.1.5 The UCLA Language Materials Project (http://www.lmp.ucla.aspx, retrieved 2007-05-16)

Teaching resources for 150 less commonly taught languages is a database of descriptions and links to materials that can be used for learning or teaching a language. No proper resources are within the database. The resources are entirely compiled by the staff members. Most of the resources listed in this database can potentially be found through search engines, but the user is at least partially assured that the resources are helpful for language learning. Problematic of course is the coverage of the database. Languages like Ami, Tayal or Tsou cannot be found in this database. We also think that any attempt to collect links to external resources risks of being overtaken by the advances of Information Retrieval. Such attempts can only be successful if they provide the possibility for collaboration (like a Wiki) or in letting people submit links to their resources, such as in OLAC.

2.1.5 More World Language Databases

There are a number of additional world language databases. A detailed discussion of these is out of the scope of this paper. In addition, the examples of world language databases discussed above suffice to motivate the design features we have selected for XNLRDF. Some of these additional world language databases, which by no means are less important or less complete are:

- UCLA Phonetics Lab Archive (http://archive.phonetics.ucla.edu/main2.htm, retrieved 2007-05-21)
- The Tower of Babel. *An Etymological Database Project* (http://smallschool.rinet.ru/main.html, retrieved 2007-05-21)
- Web Resources for African languages http://goto.glocalnet.net/maho/webresources/general.html, retrieved 2007-05-25)
- Numbers from 1 to 10 in over 5000 languages (http://www.zompist.com/numbers.shtml, retrieved 2007-05-25)
- Acientscripts.com (http://www.ancientscripts.com/ws_atoz.html, retrieved 2007-05-25)
- Geonames (http://www.geonames.de/, retrieved 2007-05-25)
- Titus (http://titus.uni-frankfurt.de/indexd.htm, retrieved 2007-05-25)
- Lowlands-L (http://www.lowlands-l.net, retrieved 2007-05-25)

2.2 Language Archives

Language Archives collect texts, dictionaries, corpora, grammars, annotated signals, interlinear texts, paradigms, field notes, linguistic descriptions, audio recordings, video recordings and other language objects. More often than not those collections focus on related languages, for example genealogically (Creek Language Archive http://www.wm.edu/linguistics/creek, retrieved 2007-05-16) or geographically (Formosan Language Archive http://formosan.sinica.edu.tw/formosan, retrieved 2007-05-16, Zeitoun 2005) related languages. A list of language archives can be found at http://www.ldc.upenn.edu/exploration/archives.html, retrieved 2007-05-15 . The examples of language archives below show their great variety. With a great number of languages in an archive, the distinction between a world language database and an archive is sometimes difficult to make.

2.2.1 Examples of Language Archives

The LACITO, *Langues et Civilisations à Tradition Orale Archive* (http://lacito.vjf.cnrs.fr/archivage/index.htm, retrieved 2007-05-16), provides free access to documents of connected, spontaneous speech, mostly in peripheral or endangered languages, recorded in their cultural context and transcribed in consultation with native speakers. At present, the archive contains some 195 documents in 43 languages.

The NTU corpus of Formosan Languages (http://corpus.linguistics.ntu.edu.tw, retrieved 2007-05-16) collects linguistic data for Formosan languages, most of which are extinct or seriously endangered. Most of them have few written records. Collecting linguistic data requires field research with audio(-visual) documentation, transcription, tagging and translation into widely used languages, here into Mandarin and English. The NTU Corpus contains three narratives in Kavalan, 22 recordings in Saysiyat, two in Amis, and three in Tsou. The database transcriptions in English and Chinese are searchable. This corpus is well-presented and richly edited. A similar example of a language archive is the *Digital Archiving Yami Language Documentation* http://yamiproject.cs.pu.edu.tw/yami, retrieved 1007-05-16 .

These projects are in many aspects typical for language documentation: A small team of researchers with limited resources focuses on one or a few languages, describe them densely with an enormous effort, and make them publicly available. For each of the language archives, however, there are the open questions whether the data format will be accessible tomorrow, whether only researchers of the concerned language can capitalize on the data, or whether the data contribute to more general issues like comparative linguistics or typological studies.

Every language archive has to re-implement a similar basic infrastructure. As a consequence, there is no or little synergy among the archives and the different archives even if they overlap in their scope of languages might be incompatible due to different annotation, transcription or encoding standards. Thus, the idea to integrate archives into a network comes quite naturally. Different kind of networks can be envisaged. Either they form an confederation where resources can be searched and exchanged from archive to archive within such a confederation (c.f. Nathan 1996), or they are even more loosely connected, via a standard set of meta-data. The latter approach is followed by OLAC and Atlantis.

2.2.2 Meta-data driven search engines

OLAC, the *Open Language Archives Community* (http://www.language-archives.org, retrieved 2007-05-16) is a consortium of 39 linguistic data archives. The project started in2000 and references more than 29,000 objects. This catalog is filled by a community which holds language resources. The description uses a set of XML-based meta-data, the OLAC metadata

(Simons & Bird 2001). Once registered, OLAC harvests the language objects and makes them accessible through an intuitive search engine enhanced by additional features like a thesaurus of alternate language names, language code searching, and keyword-in-context display in search results (Hughes & Kamat 2005).

The peculiarity of OLAC is that the data remain in their original form with the data providers, who are also responsible for the meta-data. On the one hand, this approach has the advantages that the decentralized data description is generally quite accurate (Bird & Simons 2001) and that any update of the referenced data appears in real-time by virtue of automatic data administration. The disadvantage of this approach is that "data update" might mean "data loss", because OLAC does not keep a copy of registered sources. OLAC's declared goal is to perform better than Google for linguistic resources is not met at all: excluding the references to the world language databases discussed above we find in OLAC zero entries for Ami, Bunun, Paiwan, Tayal and Tsou, in comparison many more documents in any search engine.

A software pool for peripheral European languages is the Atlantis database (http://marroc.uoc.es/atlantis/eng/database.html, retrieved 2007-05-16). It lists 1542 on-line resources, but even among peripheral languages there is a clear hierarchy of "more" and "less" used peripheral languages. In fact there are many resources for Catalan, but few for Sorbian. New entries can be registered to the database, but many links are broken .

2.4 Software Pools

A software pool is a collection of electronic resources (dictionaries, corpora, spell-checkers etc) which share functional or formal properties. According to Streiter, Scannel and Stuflesser (2007), a software pool offers researchers the possibility to register and hand over their data and thus safeguard their maintenance, updates and usage beyond what can be directly influenced by the researchers themselves. When software pools update resources for new versions they transform generally all resources of the pool, knowing that the attractiveness of the pool comes from the number of different language modules it contains. If all language modules have the same format and function and if one module can be transformed automatically, all others might be automatically transformed as well. This advantage of software pools thus comes from the uniform format and that the collection of resources has a kind of critical mass that cannot be ignored when resources are ported for the future. In addition, open software pools have an important social and sociolinguistic function:

> *... by simply making the source code and data underlying your project freely available, you enable other members of your language community to contribute to the project, or to develop their own projects based on the foundation you have provided. It is important to emphasize a relevant sociological aspect of free software here: freely available source code provides the means by which members of the community can contribute, but also provides a strong motivation, since there is often a spirit of collective ownership of the resources. We have found this to be particularly true of language processing projects, which also harness the pride many speakers have in their mother tongue.*

Streiter, Scannel and Stuflesser (2007:12)

In their discussion of software pools Streiter, Scannel and Stuflesser(2007), list a number of properties which allow to distinguish the quality of a software pool and which can certainly also be used to stipulate requirements for language archives and world language databases. These criteria guess the likelihood that data will be continued to be used, updated and ported for future applications. Most of these features seem to fit better world language database than an archive. So that we might conclude that the former more in the direction we have to follow. Archives are not polychrome, may stress the individuality over uniformity, and are maybe more paradigm dependent that world language databases.

- uniformity of the data
- maintenance by a community
- mirrored on many servers
- paradigm independence
- popularity
- variegation(many instances of a datatype)

3 Conceptual Design of XNLRDF

In this section we will discuss the basic conceptual features of XNLRDF. Most of these features follow from our analysis of related projects. Although the combination of these features in a running system is difficult to achieve, it is our utmost conviction that language documentation will go into this direction driven by requirements of portability, timeliness, quantity, quality, relevance and computability.

3.1 Structured Data

In XNLRDF we emphasize the creation of structured data. Structured data are data of a specific type, a specific range within a relatively well-defined interpretation. Structured data can be read and understood by a machine and are thus portable to the future. Unstructured data tend to be ambiguous, require interpretations and are at permanent risk to get lost if no interpretation can be assigned anymore. Actually, we think that most attempts to document a language in the form of unstructured data puts at risk the very purpose of that enterprise.

Unstructured data, one might argue, have the advantage to be more flexible and express uncertainty, the yet not defined or the undefinable. While the latter, however, are not supposed to be within a language database, the attempt to encode the former trades truth for relevance. While the indication that a language in 1988 had 30 to 70 speakers might be true, the distinction made here might not be relevant. If such distinctions are or might become relevant in the future, they can be expressed formally otherwise, e.g. with a normally distributed probability function with its maximum probability at 50 and P(29)=0 and P(71)=0. Unstructured data categories allow for true specifications, e.g. specifying the numbers of speakers as "few", missing the relevant distinction between 5, 5.000 or 50.000 speakers. In order to handle the underlying need that seem to require unstructured data, discussion or comment fields with free text input can be added to a database and actually are added to XNLRDF. The content of these comments is then to be converted, slowly by slowly first into the appropriate categories and than in structured data.

Structured data, on the other hand, can be automatically checked, to some degree, for coherence, consistency and comparability. With 20.000 languages, dialects, and language families, 100.000 writing systems, thousands of tribes and ethnic groups, hundreds of scripts, hundred of writing standards, hundreds of thousands of place names in the present and the past, the usefulness of such automatic checks can be hardly denied and XNLRDF makes extensive usage of such checks. *Coherence* refers to that property of the data which allows for a combination of elementary data of different categories in a meaningful way. To give an example, the number of inhabitants of a country and the numbers of speakers of a language in that country can be combined. If they are coherent, they yield the percentage of inhabitants of that country speaking this language. The coherence can be checked with a battery of rules, each of which implements part of the logic of the data. One possible rule, for example, might state that the number of speakers in a country cannot be higher than the number of its inhabitants. The number of speakers in several countries summed up cannot be higher than the total number of speakers of this language. The number of people and the number of speakers in a region cannot be higher than a maximal population density. The number of speakers cannot below a density threshold, to be relevant. The battery of rules thus implements a network of logical constraints which limit the range of possible data. Similar rules can be applied to other domains, e.g we can check whether the characters of a text and those associated with the text's script (Latin, Arabic, etc) are compatible, whether the IPA transliteration in interlinear texts corresponds to the list of phonemes of that language, whether script and writing standard are compatible, when writing standards and languages are compatible etc. Thus, the richer the system becomes, the more smaller the mistakes can be made.

Coherence control also indirectly the consistency, where *consistency* refers to the fact that data within one data category are comparable, e.g. that transitivity of the data can be assumed. E.g. by summing up the numbers of speakers in a country, the range for the number of speakers of successively entered languages becomes smaller and smaller and with them the range of possible mistakes. In XNLRDF, cornerstone data are entered first (population of the world, of China and India, the collapse of the Soviet Union) and frozen by superusers after inspection. These and all successive data will control what can be entered in the same and other data categories.

This summing up of inhabitants and numbers of speakers, as well as *comparability* can be achieved by having hierarchical organizations of data, i.e. tree-structures instead of lattice structures. Although this formal requirement might be unnatural, e.g. one has to decide whether a script using Latin and Chinese characters inherits its basic properties from the Latin or the Chinese script, languages per script or script class cannot be summed up and meaningfully compared, if one would allow a mixed script to inherit from two scripts. It would be neither possible to sum up the languages per script nor to compare the scripts. Having multiple inheritance with thousands of languages, hundreds of thousands of writing systems would make the data uncontrollable. Checks for re-entrance are formal in nature and easily to implement. Instead of multiple inheritance, however, one might have multiple and different inheritance principles, e.g. independent tree structures that connect the basic entities. Punctuation marks, for example might be inherited along different lines than numbers, word separators, the writing direction or the main stock of characters. This mode of a multi-layered description might thus not only maintain the requirement of having tree-like structures, but be, in addition, more precise as the different tree-structures specify different modes of relatedness. So one might let the mixed Chinese/Latin script inherit the punctuation marks from the Chinese script and the writing direction from the Latin script. This multi-layered tree structure, in addition, allows for a better, since more specific, control of the coherence and consistency for each of these layers.

3.2 Collaborative Data Creation

The collaborative creation of data, texts and software has shown to be a feasible and maybe superior model of knowledge creation. As in Open Source software projects or Wikimedia projects, professionals and non-professional experts can collaborate and accomplish gigantic amounts of labor through structured on-line collaboration. From the world language databases cited above, Wikipedia, the Rosetta Project and Language Typology make use, to different degrees of collaborative work. And only these projects do not risk to be overcome by other, faster going projects. Of course, non-collaborative projects, could take over the data of different resources, shuffle and combine them and become for some time the principle resource of data. Without a continuous elaboration of the data through collaborative work, this new database will fall out of used.

Although the question of quality has been recurrently brought up as an argument against this model (cf. Wired 2005), there is no general proof that traditional academic or industrial products are of higher quality. Even worse, many high quality resources developed by academia or industry are not available. People prefer to have their data become inaccessible on an old floppy disk over sharing and developing them with interested researchers. Quality, at the same time, has been recognized as a key notion in collaborative projects and it is tackled with technical support and the great number of collaborators. In XNLRDF, for example, coherence checks are used wherever possible to control the data input. Frequently the control goes from general data to more specific data. The feature 'noun' (has the language nouns?) has to be set to 'yes' before the feature 'n2v' (has the language a noun-to-verb derivation?) can be set to 'yes' or 'no'. Once 'n2v' is set to 'yes', 'noun' can not be reset to 'no'. 'n2v' cannot be set to 'no' as long as the example illustrating a noun-to-verb derivation in that language has not been removed. This way mistakes at a relatively high level are easily avoided as they have fare-reaching consequences, possible mistakes at the bottom end are either excluded or of limited consequences and can be corrected anyway. That this social control can work well as in Wikipedia, where an army of volunteers survey their entries like eagles their hunting areas, or as in Language Typology, where modifications are moderated.

Beside the great amount of labor that can be accomplished in collaborative projects, these projects transcend the border between producers and users. Users discuss their needs, determine what they construct and help in doing so. The data produced are needed data and those who produce the data are motivated to do so. Raising awareness and user participation, two concepts generally held important in the support of eroding languages are almost naturally supported in collaborative projects, although assuming naively that the users are connected on-line, share a common language, a common writing system etc. A world language database therefor should be localized, a work which again can be taken over partially by the users and which at the same time creates new, topical language resources in the languages localized, e.g. in the form of a linguistic terminology in a peripheral language.

3.3 Not encoded, computational Data

The *encoding* of the data is another important issue. The first level of encoding refers to the question whether or not the data are encapsulated. Only non-encapsulated data are portable, where *encapsulation* can be understood as the fact that a file can be edited safely with an editor like Vi or Wordpad, .e.g to repair distorted data. The encoding is also determined by the *compression* of data. Compressed data are very likely to be encapsulated data. In addition, compressed data are suppressed data. When data are updated from one compression format to the next generation compression format, other pieces of information might be suppressed without a compensation for the data suppressed in the previous compression format. Thus data by data will be lost, or if not updated, data will become inaccessible.

The encoding determines the *computability* of the data. What is computational, however, is a technical question. Whether a JPEG-file of a 15. century Bible is computational depends on the quality of the Optical Character Recognition (OCR). If a lot of characters cannot be recognized the data are not computational and inaccessible for any automatic processing. PDF-files containing linguistic descriptions, images of character lists or maps of language distributions are thus of limited or no use. They can be read, one by one by an interested human, but have to be transformed into other representations when data are to be combined, compared or used to check their coherence and consistency. Digitalizing must be done in a format that is fully computational, e.g. the data can be theoretically re-generated in a given standard.

As a consequence, we do not consider for the moment images, audio and video files to be primary data in XNLRDF. Images, audio and video data can be transformed, however, through human interpretation, optical character recognition through advances in speech-recognition and image-recognition into valuable resources. The images used currently in the XNLRDF browser thus serve primarily for the orientation for those who enter the data. In the near future, the images of countries and regions in XNLRDF will be replaced by the images created from a Geographic Information System. The data in XNLRDF are managed in a relational database, which is unfortunately an encapsulated format. However, frequent updates are made and stored in flat text files.

3.4 Free Data

Data produced in XNLRDF should be freely available. Many of the applications discussed above hide their data or protect them via a proprietary copyright. As discussed in Streiter, Scannel and Stuflesser 2007, this is not only counterproductive to the original purpose, i.e. making language data available, but puts the entire data at risk which, without being copied and updated at different sides might be forgotten or become inaccessible. In addition, making the data freely available is one way to motivate people to collaborate. As collaborators get neither academic credits nor money for their work, the benefit in creating free data is in creating data which are needed for the proper interests or research which is considered to be important in the future. The individual researcher thus updates the XNLRDF database and can than download the entire data stock including the personal updates plus the updates made by others. To have this immediacy, we allow users to compile their downloads. Its is our goal to make it more easy to update the data in XNLRDF and than download the data, than to download them an then update them for the current need off-line.

3.5 Tools and Utilities

In order to illustrate the usefulness of the data, and to provide some services from which the users can profit, a number of tools are arranged around the XNLRDF-interface. They are not part of the XNLRDF download, but use the data dynamically from the database. These tools are first, a concordancer for all the languages that have textual examples. Although the corpora might be small, the concordancer is a nice tools for language learners to study and understand the behavior of the basic words of a language. Second, equally fully automatically derived from the data is a spelling checker (c.f. Liu et al.2006). Here also, beginners might check their writing in Tsou before sending of an email in Tsou. The spelling checker bears also a symbolic function. It signals on the one hand, that the language is important, that it can be used for writing and that with the relatively modest effort of collecting and inserting some words, something useful can be made for a language. Finally the translation dictionaries compiled from XNLRDF are used automatically to update the linguistic backbone of Gymn@zilla, a language learning program that annotates on-line web-sites and generates exercises from them. At the current stage, however, the resources of Formosan languages in XNLRDF are not sufficiently rich in order to be taken into consideration by Gymn@zilla. Other tools can be more or less easily be derived from the data, such as a comparative grammar for two languages or an automatic comparison of the lexis of two languages.

4. XNLRDF (http://140.127.211.214/xnlrdf, retrieved 20.5.2007)

Based on Data from Ethnologue, Wikipedia and hundreds of other language-related resources, XNLRDF, the Natural Language Resource Description Framework, tries to draw different data and resources into one formal model. The highly structured data are kept in a relational database and can be downloaded under the GPL in XML. The project described in Streiter & Stuflesser (2006) started 2005.

The basic entities in the database are LANGUAGE, LOCALITY, SCRIPT, ORTHOGRAPHY, WRITING_STANDARD, PEOPLE, RELIGION, WRITING_SYSTEM, LIFE_SYSTEM, WORD, TEXT, CHARACTER, FOREIGN CHARACTER, SYMBOL, NUMBER, WORD_SEPARATOR, SENTENCE_SEPRARATOR, MEDIA FILES, NLP_RESOURCE and SOURCE. Each of these entities may have additional features, e.g. LANGUAGE has the features *a2n* (adjective to noun derivation), LOCALITY has the features *size* (the surface in square kilometers) etc. LANGUAGE, LOCALITY, SCRIPT, WRITING_STANDARD, PEOPLE and RELIGION are organized in tree structures. In addition, entities can enter relations. For example, LIFE_SYSTEM and WRITING_SYSTEM are basically characterized by LANGUAGE, LOCALITY and a *time period*. WRITING_SYSTEM, in addition, has a SCRIPT, a WRITING_STANDARD and an ORTHOGRAPHY (Streiter and Stuflesser 2005). A LIFE_SYSTEM has a mode. SCRIPT and WRITING_STANDARD are interlinked as some scripts can only be used for a limited number of writing standards. TEXT, WORD, CHARACTER, NUMBER, FOREIGN CHARACTER and SYMBOL are assigned to WRITING_SYSTEM. WRITING_SYSTEM can be linked to WRITING_SYSTEM as 'transliteration of' and LIFE_SYSTEM and LIFE_SYSTEM may be interlinked as 'transcription'. LIFE_SYSTEM can be linked to audio and video-files. TEXT is linked to WRITING SYSTEM. Grammatical examples and glossed interlinear texts in LANGUAGE are related to a writing system of a language and are thus also texts within a writing system. Language examples can be given in any writing system of language. Texts may be related among each others. When they are related within one writing system they represent a paraphrase, when related within one language they represent a transliteration and when they cross language borders they represent translation. Texts may also be linked to media files of the same language and thus represent transcriptions, or when linked to media files of other languages they represent transcriptions of translations.

Below we present the principal data structure. The first row, for example, reads like follows. The entity is a language. The language has an internal identifier, a data for the birth of the language and a date for the death of the language. The language family is a pointer to another language. The language may follow another language, like Classical French followed Middle French and precede another language like Classical French precedes Modern French. A great number of features describing the language follow. The default writing system can be reset. This parameter specifies the writing system in which the currently visible examples for grammatical structures are written.

[LANGUAGE: id, valid_from, valid_to, language_family -> **LANGUAGE**.id, follows -> **LANGUAGE**.id, precedes -> **LANGUAGE**.id, default_writing_system, prepositions, postpositions, circumpositions, p_n, n_p, card_n, n_card, n_gen, title_name, name_title, gen_n, aux_v_main, v_aux_main, v_aux_sub, aux_v_sub, c_s, s_v, svo, ovs, vso,vos, sov, osv, svo_main, ovs_main, vso_main, vos_main, sov_main, osv_main, clitics, serial_verbs, pronom_clitics, wh_movement, wh_in_situ, wh_in_situ_optional, multiple_wh_movement, nominative_accusative, passive, ergative_absolutive, n_num, a_num,d_num, n_poss, num_n, num_a, num_d, poss_n, genitive, adv,tonal, vowels_nb, nasal_vowels_nb, consonants_nb, diphtongs_nb, phonem_vowel_length, cv, v, cvc, ccvc, ccvcc, ccvccc, cvcc, cvccc, ccvccc, cvcccc, ccvcccc, cccvccc, ccv, vc, vcc, vowel_harmony, consonant_harmony, vowels, consonants, nasal_vowels, diphthongs, phonemic_stress, word_final_stress, word_prefinal_stress, word_initial_stress, root_initial_stress, root_final_stress, root_prefinal_stress, phonemic_intonation_nb, n, verb, a,num, d, poss, a2n, a2v, v2n, n2n, n2a, n2v, a2adv, adv2a, noun_cases, composition, prefix, infix, suffix, circumfix, phonemic_intonation_nb,nominative, coordinating_connective, enclitic, proclitic, wh_movement_optional, singular, plural, large_plural, dual, trial, several, multal, paucal, small_paucal]

The different names of a language are stored as language name, the name is given in a written form which belongs to a writing system and might have additional temporal constraints.

[LANGUAGE_NAME: id, name, language_id -> **LANGUAGE**.id, valid_from, valid_to, default, deprecated, writing_id -> **WRITING_SYSTEM**.id , ...]

The life system is a language in its spoken, signed, whistled, hummed or sung form. A life system is mainly characterized by the language, the locality (geographical and administrative) and a time period. A life system might be restricted to a specific mode (singing, speaking, whistling) and used only by one gender. All other databases ignore the distinction between a languages and a life system, describing whistled forms of speech not as another life system but as another language. However, whistled, summed, drummed, sung or spoken, they more often than not are the same languages. In addition, features assigned to the life system are not features of a language. A language cannot be restricted to a certain gender or a certain region, but life systems can. A language as a means of socially determined cognitive categorization cannot be forbidden, a life system can.

[LIFE_SYSTEM: id, valid_from, valid_to, precedes -> **LIFE_SYSTEM**.id, follows -> **LIFE_SYSTEM**.id, language_id -> **LANGUAGE**.id, geo_id -> **GEOGRAPHICAL_REGION**.id, admin_id -> **ADMINISTRATIVE_REGION**.id, mode, gender]

Currently XNLRDF has only one type of locality. This will be split in the near future into geographical regions and administrative regions. The geographical data will be stored here,

[GEOGRAPHICAL_REGION: id, size, coordinates, ...]

[GEOGRAPHICAL_REGION_NAME: id, name, geographical_region_id -> **GEOGRAPHICAL_REGION**.id, valid_from, valid_to, default, deprecated, writing_id -> **WRITING_SYSTEM**.id , ...]

Temporal limitations are stored in the administrative_region.

[ADMINISTRATIVE_REGION: id, valid_from, valid_to**, ...]**

[ADMINISTRATIVE_REGION_NAME: id, name, administrative_region_id -> **ADMINISTRATIVE_REGION**.id, valid_from, valid_to, default, deprecated, writing_id -> **WRITING_SYSTEM**.id, **...]**

Geographical regions stand among each other in relations such as **OVERLAP, BORDER, INCLUDE** and **NOT_CONNECTED**. Administrative_regions are organized in the hierarchical **CONTROL** relation. Administrative and geographical regions are connected through the relations **ADMINISTERS**, **PARTIALLY ADMINISTERS** and **BORDERS**.

A writing system is specified through a language, the time, the place, the script, the standard, the orthography and the default_encoding. Writing systems stand in a similar relation to a language as the life systems do, the distinction though between language and writing system are more obvious.

[WRITING_SYSTEM: id, valid_from, valid_to, precedes -> **WRITING_SYSTEM**.id, follows -> **WRITING_SYSTEM**.id, language_id -> **LANGUAGE**.id, geo_id -> **GEOGRAPHICAL_REGION**.id, admin_id -> **ADMINISTRATIVE_REGION**.id,script_id -> **SCRIPT**.id, orthography_id -> **ORTHOGRAPHY**.id, standard_id -> **STANDARD**.id, default_encoding -> **ENCODING**.id, transcript -> **LIFE_SYSTEM**.id, corpus_types, corpus_tokens, ...]

The characters of one writing system might be mapped onto the characters of another writing system when then two systems stand in a relation of transliteration.

[TRANSLITERATES: id, transliterates -> **WRITING_SYSTEM**.id, transliterated ->

WRITING_SYSTEM.id, replacing_string, replaced_string]

A script describe a general set of symbols which is than mapped, through a standard and an orthography to words, syllables or phonemes.

[**SCRIPT**: id, valid_from, valid_to, unicode_covered, derived_from -> **SCRIPT**.id, precedes -> **SCRIPT.id**, follows -> **SCRIPT**.id, script_type, iso15924, iso15924_num ...]

Script may have different names and are listed in SCRIPT_NAME.

[**SCRIPT_NAME**: id, name, script_id -> **SCRIPT**.id, valid_from, valid_to, default, deprecated, ...]

A standard defines the principles according to which the graphemes and phonemes are mapped (if at all). A standard may be specific to a group of languages or a group of scripts, identified by their top-nodes.

[**STANDARD**: id, valid_from, valid_to, substandard_of -> **STANDARD**.id, script_id -> **SCRIPT**.id, language_id -> **LANGUAGE**.id, ...]

A standard can have many different names.

[**STANDARD_NAME**: id, name, standard_id -> **STANDARD**.id, valid_from, valid_to, default, deprecated, ...]

User Participation in World Language Databases

An orthography further specifies the standard and does so by referring to morphemes, words or phrases.

[**ORTHOGRAPHY**: id, valid_from, valid_to, script_id -> **SCRIPT**.id, language_id -> **LANGUAGE**.id, ...]

An orthography can have many different names. As most orthographies are just called 'old' or 'new', somewhat artificial names have to be used here.

[**ORPTHOGRAPHY_NAME**: id, name, encoding_id -> **ORTHOGRAPHY**.id, valid_from, valid_to, default, deprecated, ...]

The entities **CHARACTERS**, **FOREIGN_CHARACTERS**, **SYMBOLS**, **NUMBERS**, **WORD_SEPARATORS**, **PHRASE_SEPARATORS**, **WORDS**, **TEXTS** are all linked to **WRITING_SYSTEM**. and **WORDS** and **TEXTS** must be within the scope of the **CHARACTERS**. A subset of the entities in **TEXTS** are examples of grammatical, morphological or phonological structures for a language. The writing system, in which the examples are given can be set and reset in the language (default_writing_system). The writing system in which the translations of the example sentences are to be shown can be set in the same way.

[**TEXTS**: id, writing_id -> **WRITING_SYSTEM**.id, parallel_text_id, texts, url, form_id -> **FORM**.id, subject_id -> **SUBJECT**.id, publication_date, license -> **LICENSE**.id, copyright_holder -> **COPYRIGHT_HOLDER**.id, ...]

The **FORM**, **SUBJECT**, **LICENSE**, **COPYRIGHT_HOLDER** are than specified by specific tables, each of them, again, link to a table containing the respective names. Identical parallel texts ids mark translations (different language), transliterations (same language, different writing system) and transcriptions, i.e. when the same parallel text id is shared between a text and a media_file.

[**MEDIA_FILE**: id, life_system_id -> **LIFE_SYSTEM**.id, gender, age, media_url, cache_file, format_id -> **FORMAT**.id, kilobytes, minutes, parallel_text_id, publication_date, license -> **LICENSE**.id, copyright_holder -> **COPYRIGHT_HOLDER**.id, form_id -> **FORM**.id, subject_id -> **SUBJECT**.id, ...]

People are grouped into PEOPLE, where the criteria for the classification, in doubt, should be that the people apply to themselves.

[**PEOPLE**: id, valid_from, valid_to, subgroup_of -> **PEOPLE**.id, ...]

[PEOPLE_NAME: id, name, people_id -> **PEOPLE**.id, valid_from, valid_to, default, deprecated, ...]

Which bunch of people lives where? How many of them in which year? Are all age groups present? These questions are answered in the table LIVE_IN.

[LIVE_IN: id, people_id -> **PEOPLE**,id, geo_id -> **GEOGRAPHICAL_REGION**.id, admin_id -> **ADMINISTRATIVE_REGION**, number, census_date, age_from, age_to, ...]

FIRST_LANGUAGE_SPEAKERS and SECOND_LANGUAGE SPEAKERS describe which PEOPLE speak where and in which year which LANGUAGE. Do all age groups speak the language?

[FIRST_LANGUAGE_SPEAKERS: id, people_id -> **PEOPLE**,id, life_system_id -> **LIFE_SYSTEM**.id, number, census_date, age_from, age_to, ...]

[SECOND_LANGUAGE_SPEAKERS: id, people_id -> **PEOPLE**,id, life_system_id -> **LIFE_SYSTEM**.id, number, census_date, age_from, age_to, ...]

The output data for the download are organized in structures similar to the data structure shown here. One download file might correspond to one, two or tree of these table, in the latter case linked together through the corresponding Ids.

5. The impact of XNLRDF

In this section we will try to identify the potential impact of XNLRDF on the Austronesian languages and particularly the Austronesian languages in Taiwan. We claim that Taiwan through its political, institutional and scientific situation offers an opportunity for the development of XNLRDF and to create through it a dynamic environment for the collaborative creation of horizontal and vertical, diachronic and synchronic, qualitative and quantitative data which can diffuse into nearby related fields and areas.

Taiwan and Austronesian languages in Taiwan offer probably the best geographical, political, institutional and scientific environment for developing linguistic data through a collaborative tool like XNLRDF. First of all, these languages form a relatively homogeneous geographic and linguistic entity. Moreover these languages are all in a middle term endangered and have approximately the same low status in the society (c.f. Ferguson 1959). Second, the languages are accessible given their limited number, the limited number of speakers, their geographic location and the manageable number of administrative bodies in charge. The situation contrasts thus sharply with the environment we find in Indonesia , Malaysia , Philippines or Madagascar where languages have various official, vehicular or regional status (Raillon 1999). Experiencing the same threat of extinction in a Mandarin-dominated Taiwan, the different language groups developed a long-standing solidarity. The political and institutional environment is also excellent in Taiwan in promoting Taiwanese cultures including its aboriginal cultures. Research on aboriginal languages and cultures is encouraged by the authorities and supported by many institutions. Furthermore, the research for these Austronesian languages in Taiwan is excellent in terms of initial formation, quality of fieldworks, relevance of scientific publications and the involvement of informants in preservation projects.

Such community members involved in the preservation, promotion and documentation of endangered Austronesian languages, if providing data, could demonstrate the relevance of such a database, its capacity to associate specialist and non expert informants and its diffusion by capillarity to other Austronesian language areas. Such a data stock will allow to develop simultaneously in various directions, horizontal and vertical, diachronic and synchronic qualitative and quantitative.

The horizontal dimension is the storing of data of Austronesian languages from other parts of the world, geographically or linguistically close to Austronesian languages in Taiwan, according to the interest and resources of any data provider. The vertical data describe extinct languages such as Takupulan/Takupuyanu or Siraya. These data on extinct languages would also be completed in a diachronic direction of a language, or in synchrony direction through the possibility of creating new languages or within the same languages new life systems, e.g. when the same language undergoes significant political or socio-cultural changes.

The principle of collaborative association of specialists and non specialist as could be created through a database like XNLRDF would allow to generate more up-to-date, more accurate and more relevant data and to share knowledge about linguistic resources and the languages they both are interested in.

6. Difficulties

There are a number of problems we encountered in our work we consider worth discussing here.

First, it seems to be difficult to motivate people to work on a such a huge database. Any single piece of information entered seems to be insignificant in comparison to the global data space. In addition, it is difficult for a user to identify emotionally with the database. A more localized database adorned by cultural markers, focusing on a certain subset of the data would create a cozy niche in which users might be willing to spend time on tedious data insertion. It will be thus one of our most important aims in the near future to develop localized skins and filter to the database. A very first skin defined as

```
[SKIN: background_image, cultural_item_image, cultural_sound_file, geo_id -> GEOGRAPHICAL_REGION.ID, language_id -> LANGUAGE.ID],
```

as recently implemented in XNLRDF, does not seem to be sufficient for this purpose. Skins have to more different, more localized, provide localized search options and provide more items which allow to identify with. Below we will show what this might look like.

It is neither easy for the user to understand the purpose of the database. Since most web-sites and language archives show nicely presented data to an information-seeking user, this is what people expect XNLRDF to do. However, the Web-interface of XNLRDF is, at best, an interface for data insertion, since the focus of XNLRDF is on data and making data available as free downloads. Of course, future work will also try to make the interface more fashionable, but the main difference between this and other databases will continue to exist.

In addition to these two first strategies, another important step would be to offer the common interface not exclusively in English. Although being the global language for research, English can be an obstacle to many potential informant who are not familiar with it or its writing. Using localization techniques, this interface could show up in the main international languages as e.g. the official languages at the UN with the possibility to add new languages and attach them per default to a skin.

Another point is that people who use XNLDRF are not familiar with linguistic notions and adding simple translations of the definitions of the data categories in the localized interface languages, or either using illustrative images such as used in Pehlivanova and Lebedeva (2006) or in other cognitively inspired grammars, e.g. Schwarze (1988), would smoothen the interaction of informants with the database. The informant then could get a minimal introduction into linguistic concepts and reasoning in exchange for data they provide. This might go so far that an informant might become a specialist. XNLRDF could then be used as in-class tool in linguistic courses also.

A fourth point is that most of these extra specialist informants are not aware of most syntactic and morphological features but experience the variation as changing sounds or words and only these they could register or write down. But these variations, as important as they are for a linguist, don't fit the immediate needs or interests of the local informant. Their language experience is direct and it would be a possible attraction for them to fill in data for or get data from close but distinct language communities as e.g. in Tsou between the Tfuya and Tapangʉ, or these and Duhtu or Kanakavu. The system then could list the communities and differences, creating at the same time the sense of community. As a kind of common reference across language boundaries and as a close substitute for immediacy, images can be used to trigger the insertion of words or to highlight the difference of word meanings. Similar to *dicts.info* (e.g. http://www.dicts.info/uda2.php?k=128, retrieved 2007-05-25), where a multilingual lexicon is build with images and English definition as interlingual glue, a localized set of images of natural object (sun, moon, dear, mountain, etc.) and of local, cultural objects might be at the center of such an immediate contrastive analysis. The set of images might be registered with the skin by the informants for a language and thus spread over all languages connected by the skin.

Through the above elaborations it becomes obvious that the gigantic data downloads of XNLRDF are interesting for professional researchers, but quite useless for an average user, who at best would work, but for what reason, with a spreadsheet program. Therefore, smaller, more witty downloads and dynamic services, also interesting for average users should be envisaged to compensate for the users' efforts. Creating dynamically a PDF-file of a contrastive image dictionary for free download would be one of the many options. Another option is to provide text and audio-files in such a way that both text and sound can be seen and heard in parallel on an Ipod.

Another, completely different problem related to the work on XNLRDF relates to copyright issues. As XNLRDF tries to provide data under the GNU Public License, only data that can be distributed under this license can be inserted into the database. For atomic data, this is not a real problem, as no-one has the copyright, for example, on the penultimate stress in Tsou. In addition, we believe that no linguist would object to a reproduction of his or her glossed interlinear texts, if correctly cited. Text-files and audio-files are a more serious problem. Copyrighted texts can be upload in XNLRDF, but not distributed as such. All other information that can be derived from the texts and of which the author of the text cannot claim the copyright can be distributed however. This means, for example, that word-list, n-grams, frequencies, probabilities etc. can be extracted and distributed. Therefor for each uploaded text we specify the copyright as far as possible, but then text files disappear from the screen and show up, e.g. as frequency lists. Audio files are not included in the XNLRDF download, as the download is entirely in XML or flat text files. They can be downloaded however from the cache in which we have the files, similar to the cache of Google's files. In this case, downloading the files from the original site or XNLRDF is essentially the same.

7. Summary and Discussion

In this paper we presented a system which tries to provide free large-scale in-depth information on the languages of the world. We have argued that the approaches followed by parallel projects with a similar goal don't seem to be promising to us. Any attempt, for example, to index external resources has to compete with powerful commercial Information Retrieval systems, has to handle broken links, and does not improve the interoperability or portability of the data themselves. Language archives have to struggle with the interoperability of the data and frequently severely restrict the usage of the data through the type of copyright they impose. In addition, they offer no transfer strategies of how to port the system to other language groups in other areas of the world.

The points for structured, computational and portable data has also been discussed extensively. Data should be structured to remove ambiguity in the data as much as possible. Data should be computational in order not to require the intervention of an interpreting program before accessing the data. Data should also be portable, thus neither compressed nor encapsulated. Finally, data should be free. Any attempt to gain an advantage by making the data not freely available is detrimental to the destiny of the data, as shown in cf. Streiter, Scannel and Stuflesser (2007).

In addition, although automatic transformation and acquisition of data is done in the background, the main data have to be entered manually. Since XNLRDF is formal and much more specific that other databases or formal descriptions (e.g. we do not only work with notions like 'English' or 'French' but with hundreds of writing and life systems for each of them), the insertion of data requires to identify, if needed, the right life system or the right speech system. Thus, before data can be inserted, they have to be interpreted by humans and the ambiguity in the data has to be resolved. This enormous labor can only be tackled with a large army of volunteers which is guided by a good interface and controlled by a strong database backbone. Users are compensated for their encoding effort by services or downloads which include their recent personal modifications. The difficulties in attracting volunteers has been discusses as well as some possible ways to increase the attractiveness of the database and make it at the same time more useful.

In addition, we discussed questions as how the quality of the data can be assured. There are a number of answers to the problem, only the first one being already implemented. First, the database backbone controls what is logically possible given the current data set. Second, answers should not be seen as answers to 'yes' or 'no' questions, but as a distribution of 'yes' and 'no' answers given by different users and supported by examples of different degrees of elaboration. The assumed value then, e.g. whether or not Tsou has WH-movement, might be calculated dynamically out of these data. More important however, is to keep the data, as the judgment and its implications might change in the future. In addition, we discussed ways how the individual users can be judged. Do users contradict themselves or give impossible answers? Are their inputs frequently refused by the system? How many successful updates has the user already made? This evaluation of the user will then effect the impact of inserted data.

To sum up, currently topical items, such as domain modeling, user modeling, social information retrieval, on-line collaboration, as well as a close investigation of the linguistic needs of language groups will bring fruitful insights to this endeavor to build up large-scale formal data of the languages of the world. This is a tedious and time consuming enterprise. But there are no alternatives.

The XNLRDF project set out to support people involved in the research and support of Austronesian languages. It provides a tool with which the interested community can bundle and direct its dynamics by working on the database with data on Austronesian languages in Taiwan, by discussing the categories with colleagues and students and by suggesting better ways to capture relevant information. In that way, this community would proof it's vitality, solidarity and ability to cooperate on the preservation, promotion and documentation on Austronesian languages in Taiwan and propose a model of such actions for similar fields like Polynesia or New Caledonia through different skins.

References

Bird, Steven and Gary Simons. 2001. OLAC Overview. http://www.language-archives.org/docs/overview.html, retrieved 2007-05-16.

Bow, Cathy, Baden Hughes and Steven Bird. 2003. Towards a General Model of Interlinear Text, Workshop on *Digitizing & Annotating Texts & Field Recordings*. LSA Institute, Michigan State University, July 11-13, 2003.

Eisenlohr, Patrick. 2004, Language Revitalization and New Technologies: Cultures of Electronic Mediation and the Refiguring of Communities. *Annual Review of Anthropology 3:21-45.*

Farrer, Scott and Terence Langedoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7(3):97-100.

Ferguson, Charles. 1959. Diglossia. *Word* 15:325-340.

Gartner Bettina and Oliver Streiter. The Whistled Languages of La Gomera (Spain), Antia (Greece) and Kuskö (Turkey) - state of research and open questions. In: Abel, Andrea, Mathias Stuflesser and Magdalena Putz (Eds.) (2006): *Mehrsprachigkeit in Europa: Erfahrungen, Bedürfnisse, Gute Praxis. Tagungsband. - Plurilinguismo in Europa: esperiene, esigenze, buone pratiche. Atti del convegno. - Multilingualism across Europe: Findings, Needs, Best Practices. Proceedings*. August 24-26, 2006, Bolzano/Bozen. Bozen: Eurac. http://140.127.211.214/publs/files/pfeifsprachen.pdf, retrieved 2007-05-24.

Good, Jeff and Calvin Hendryx-Parker. 2006. Modeling Contested Categorization in Linguistic Databases. in: *2006 E-MELD Workshop on Digital Language Documentation, Tools and Standards, the State of the Art*. Michigan State University in East Lansing, Michigan, June 20-22, 2006.

Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernhard Comrie. 2005. *The World Atlas of Language Structures.* Oxford University Press, Oxford.

Hughes, Baden and Amol Kamat. 2005. A Metadata Search Engine for Digital Language Archives. DLib Magazine 11(2), February 2005, http://www.dlib.org/dlib/february05/hughes/02hughes.html, retrieved 2007-05-13.

Pehlivanova, K.I and M.N. Lebedeva. 2005. *Grammatika russkogo jazyka v illjustratcijah*, Russkij Jazyk, Moscow.

Schwarze, Christoph. 1988. *Grammatik der italienischen Sprache*, Narr, Tübingen.

Simons, Gary and Steven Bird. 2001. OLAC Meta-data Set. http://www.language-archives.org/OLAC/olacms.html, retrieved 2007-05-25.

Simons, Gary F., Brian Fitzsimons, D. Terence Langendoen, William D. Lewis, Scott O. Farrar, Alexis Lanham, Ruby Basham, Hector Gonzalez. 2004. A Model for Interoperability: XML Documents as an RDF Database, in: *EMELD Workshop on Databases*, Detroit 2004.

Lewis, W. D.. 2006. ODIN: A Model for Adapting and Enriching Legacy Infrastructure, in *Proceedings of the e-Humanities Workshop*, held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing', Amsterdam, The Netherlands.

Liu Demi Yi-Chien, Simon Chun-Feng Su, Laurel Yu-Hsuan Lai, Ellie Hsiao-Yun Sung, I-ling Hsu, Sibyl Yin-Chi Hsieh and Oliver Streiter. 2006. From Corpora to Spell Checkers: First Steps in Building an Infrastructure for the Collaborative Development of African Language Resources. *Strategies for developing machine translation for minority languages. LREC Workshop Genova*, Italy, 23 May 2006. http://140.127.211.214/publs/files/liu_etal.pdf, retrieved 2007-05-25.

Mummendey, Hans Dieter. 1995. *Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*, Hogrefe, Verlag für Psychologie.

Nathan, David. 2006. Foundation of a Federation of Archives, in: International Workshop *Towards a Research Infrastructure for Language Resources*, LREC Workshop, Genova, Italy, May 22, 2006.

Raillon, François. 1999. *Indonésie, la Rénovation d'un Archipel*, Paris: Belin, Collection Asie plurielle.

Streiter, Oliver. 2007. The Role of Geographic Information Systems (GIS) in Linguistic Research. International Conference on *Multi Development and Application of Language and Linguistics (MDALL)* National Cheng Kung University, Tainan City, Taiwan, May 31 - June 1, 2007. http://140.127.211.214/publs/files/gis.pdf, retrieved 2007-05-24.

Streiter, Oliver and Kevin P. Scannell and Mathias Stuflesser. to appear 2007. Implementing NLP Projects for Peripheral Languages: Instructions for Funding Bodies, Strategies for Developers, To appear in: *International Journal of Machine Translation*. http://140.127.211.214/publs/files/mt_min_10.pdf, retrieved 2007-05-24.

Streiter, Oliver and Mathias Stuflesser. 2006. Design Features for the Collection and Distribution of Basic NLP-Resources for the World's Writing Systems. In: International Workshop *Towards a Research Infrastructure for Language Resources*, LREC Workshop, Genova, Italy, May 22, 2006. http://140.127.211.214/publs/files/st_st3.pdf, retrieved 2007-05-24.

Tung, Tung-ho. 1964. *A descriptive study of the Tsou language*. Formosa, Taipei.

Wired. 2005. Wikipedia, Britannica: A Toss-up. http://www.wired.com/culture/lifestyle/news/2005/12/69844, retrieved 2007-05-23.

Zeitoun, Elizabeth. 2005. Préservation et revitalisation des langues formosanes : le projet "Formosan Language Digital Archive", http://www.cefc.com.hk/taipei_uk/seminaire.php?idsem=16, retrieved 2007-05-22.

The Authors

Oliver Streiter, National University of Kaohsiung, ostreiter@nuk.edu.tw
Chun-feng Su, National Sun Yat-Sen University, ahome0304@yahoo.com.tw
Leonhard Voltmer, European Academy Bolzano/Bozen, lvoltmer@eurac.edu
Yoann Goudin, National Cheng Kung University, yoanngoudin@yahoo.fr