LANGUAGE ARCHIVES NEWSLETTER

No. 8
Oct 2006

In This Number:

Editorial

Endangered Languages

Documenting Plant and Animal Names 2
Gail Coelho

Archiving

IMDI Metadata Field Usage at MPI 6
Alex Klassmann, Freddy Offenga, Daan Broeder, Roman Skiba

Unique Resource Identifiers 8
Daan Broeder, Eric Auer, Peter Wittenburg

News in Brief

Language Archives Inspire Publishing Donation 9
Pauline Edwards

DAM-LR Training and Workshop, Lund, Jan 2006 9
Sven Strömqvist

MPI Archive Services Expanded to New Audiences 10
Paul Trilsbeek, Jacquelijn Ringersma, Peter Wittenburg

Announcements

Synpathy Syntax Editor 10

DELAMAN IV at ELAR 11

Language Archives Newsletter. Web: http://www.mpi.nl/LAN/. Mail: LAN@mpi.nl. ISSN: 1573-4315.
Editors: David Nathan, Paul Trilsbeek, Marcus Uneson. Layout: Marcus Uneson.

Editorial

This issue of LAN features an article about documenting plant names of endangered languages. Gail Coelho shows how communities' knowledge of their environment is inseparable from their language, thus linking languages to the key issues of our time: ecologies, landscapes and climate. There is a technological link as well: more and more fieldworkers are using photography and video, aided by the increasing quality and convenience of digital equipment. She provides valuable advice about integrating photography into research, from avoiding thorns when photographing to collaborating with botanists.

Other items in the issue address our everyday concerns with preservation, including metadata, resource identifiers, and training.

One of LAN's aims is to foster the relationship between language documenters and archives, by bringing practitioners together and increasing awareness among linguists of the value of storing their data in specialised archives.

The editors would like to apologise for the delay in producing this issue, which has been caused by illness and organisational matters at one of our sites. In addition, we welcome Paul Trilsbeek to the team as a new editor, and thank Roman Skiba for his extremely kind collaboration as former editor.

David Nathan, Paul Trilsbeek, Marcus Uneson

Suggestions and contributions welcomed at:
LAN@mpi.nl

Next deadline for copy:
November 25, 2006

Endangered Languages

Documenting Plant and Animal Names

Gail Coelho
SOAS, London

Introduction

Many of those communities who speak endangered languages live in rural areas and lead traditional lifestyles which are directly dependent on the utilisation of local natural resources; for example, farmers, fishing communities, and hunter–gatherers who live in forested areas. These communities have a deep knowledge of their environment: knowledge of what items in their environment are useful and what are not, what resources are available, and when and how these resources can be optimally used. Their knowledge of their local environment becomes, inevitably, encoded in languages not only in the form of names for plants, animals, weather conditions, and landscape features, but also as grammatical features such as classifiers, spatial terms, etc. A community's understanding of its natural environment is reflected in the way in which aspects of its environment are categorised through lexical or grammatical items in the language. In addition, a community's cultural perspective on its natural environment is revealed in the portrayal of plants, animals, and other environmental features in verbal genres such as riddles, proverbs, folktales, creation myths, and ritual speech forms. Documentation of such environmental knowledge is important because language loss does not occur independently of loss of cultural traditions and knowledge. Communities generally regret the loss of cultural traditions and knowledge systems as much as they regret the loss of their ethnic language; therefore, a good way in which language documenters can contribute to the long-term preservation of a language is to gather records about the language in a way that maximises the amount of cultural information contained in these records.

One obvious way that languages encode knowledge of the environment is in vocabulary for plants and animals, and parts thereof, as well as vocabulary for gathering and preparing these items for use in food, tool-making, etc. In gathering names for flora and fauna, the linguist should keep in mind the usefulness of these records for a multidisciplinary audience and the fact that this information is a valuable part of the cultural heritage of a people, which must be recorded for posterity. With this purpose in mind, it is important that names of plants or animals be glossed with accurate botanical or zoological names. Detailed records of this kind enable later generations of community members to identify the species that the name refers to. Further, given that many species within the environment are themselves, in many cases, endangered, well-documented information about plants and animals serves as a record about the species, even after the species becomes extinct.

Matching native language names to the correct biological taxonomical name is, in most cases, not a task that a linguist can undertake alone. Linguists need to collaborate with biologists for this work, especially in areas rich in biodiversity where indigenous communities can have names for hundreds of species whose biological names appear only in highly technical books on plant and animal taxonomy. It requires expertise in taxonomy to recognise just what details of the plant or animal are relevant in identifying the appropriate botanical or zoological name. To complicate matters further, sometimes a single plant name in an indigenous language can refer to a range of species. Or, several distinct native names may be used to distinguish plants that, to a botanist, are merely minor variants of a single plant species. Expert understanding of how to recognise differences between plant species is necessary to detect whether the consultant consistently uses a single name for the same species, different names for a single species, or a single name for more than one species that bear a close resemblance to each other. In my own documentation of plant and animal names in Betta Kurumba (described below), I collaborated with ecologists who were studying the local flora and fauna of the region in which I was working. This article describes and evaluates the techniques I used to elicit ethnobiological names.

The language and its speakers

My field research was carried out with speakers of Betta Kurumba, a south Dravidian language. The Betta Kurumbas are a community of approximately 6000 people who live in the states of Karnataka, Kerala, and Tamil Nadu. Their territory lies primarily in the Western Ghats, a mountain range that runs along the western coast of the Indian peninsula. They also live in a section of the Nilgiri Mountains, which branch out from the Western Ghats in an eastward direction. My field site lay within the Mudumalai wildlife sanctuary, which, as shown in Figure 1, is located in Tamil Nadu, close to the point at which this state intersects with Karnataka and Kerala.

The Western Ghats and the Nilgiris have one of the highest rates of biodiversity in the world. Because the annual monsoon rainfall varies across the area, it can be divided into four main vegetation zones, listed in descending order of annual rainfall: semi-evergreen, moist-deciduous, dry-deciduous, and dry thorn (Figure 2). During my fieldwork, I chose one or two plots from each vegetation zone in which to walk with my consultant as I collected plant names. Due to practical constraints, these plots were situated outside the Mudumalai sanctuary; each plot is marked with a square in the appropriate colour showing the vegetation type of the plot.

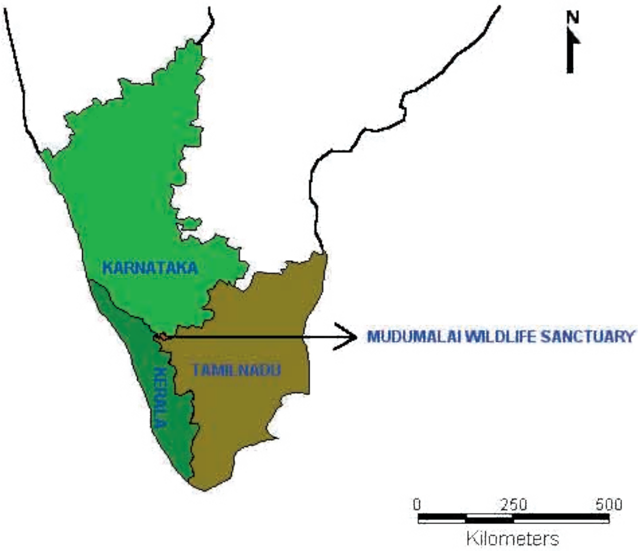


Figure 1: Location of the field site (southern India)

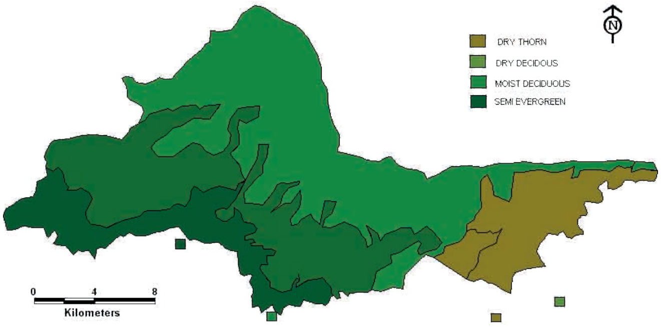


Figure 2: Vegetation types in Mudumalai and locations visited

The Betta Kurumbas are one of 16 ethnic groups indigenous to the mountain ranges of the Nilgiris and Western Ghats. They were traditionally a forest-gatherer community, but today they have limited access to forest land because most of their traditional territories have either been taken over by private landowners who have set up tea and coffee plantations or have become protected forest areas. Although Betta Kurumbas continue to live in or close to the forest and continue to use forest resources, their access to these resources is heavily restricted. Their intimate knowledge of the forest and their hundreds of names for plants and animals are in danger of being lost as younger generations find less opportunity or desire to use traditional methods. My research, therefore, is an attempt to document their knowledge of plants and animals and contribute towards preventing this loss.

Collecting plant images and names

At first, I used books containing pictures of plants in India, since books with a more regional focus on the Nilgiris, the Western Ghats, or even southern India

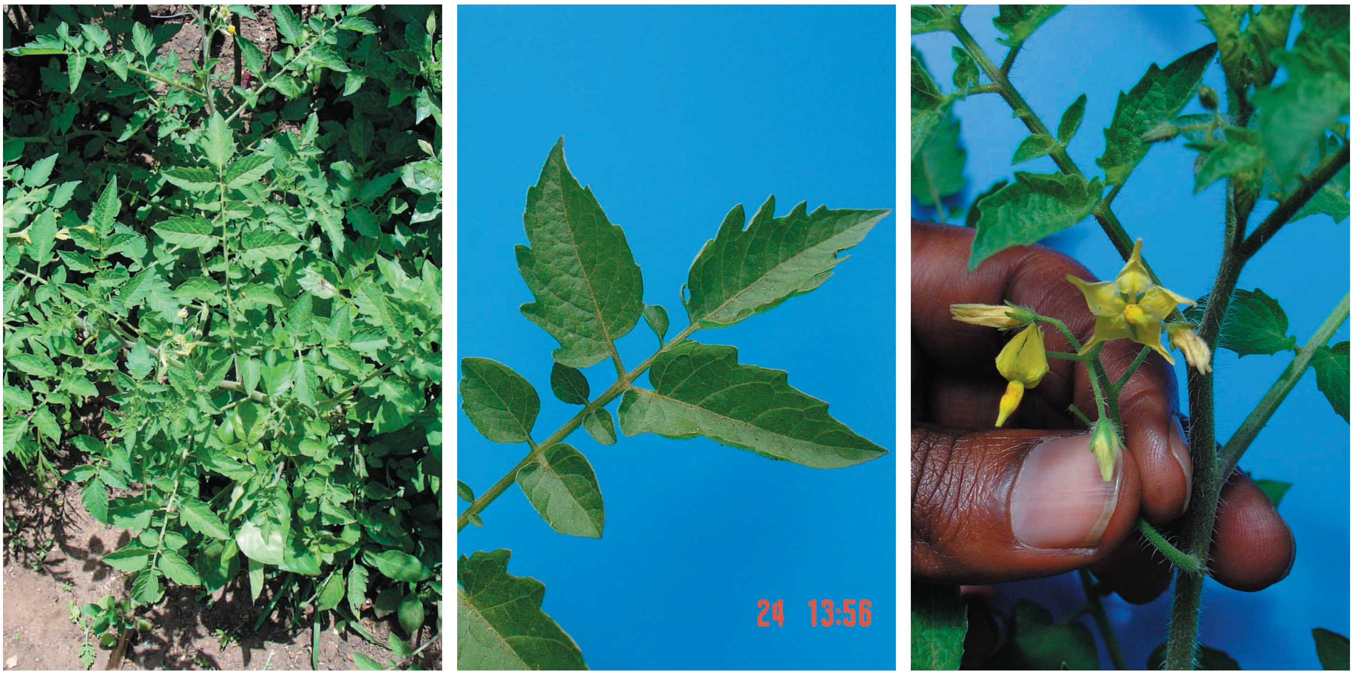


Figure 3a-c: Pandlpëyriyë (Lycopersicum esculentum)*. Whole plant (left), leaves (middle), flower (right) (cropped by the editor)*

were not available. This method was, however, very unsatisfactory for several reasons: First, only a small fraction of Nilgiri plants were actually represented in these books. Second, books tend to have a single picture per plant – a single picture does not contain all the information that consultants rely on to identify plants. They may not, for example, show flowers, fruits, or a clear picture of the background, all in a single picture. Third, a consultant might mistake pictures of nonnative plants for native plants that grow in their own environment; pictures of similar plants can be confusing for consultants who are used to relying on contextual clues when recognising objects in their environment. Fourth, with a real fruit or flower, the consultant can turn the item around, or break it up to examine it in more detail; a luxury that is not available with pictures.

Local research institutes such as the Indian Institute of Science in Bangalore have created herbaria housing labeled plant samples of all plants found in the Nilgiris. I could have tried eliciting plant names by taking my consultants to look at these herbaria; however, I anticipated that the absence of contextual clues derived from seeing the plant in its real-world environment would again pose problems for my consultants. Therefore, the method I decided on was to trek in the countryside with my consultants, taking photographs of the plants for which my consultants provided native language names. I would then show these photographs to a botanist at the Indian Institute of Science, who would consult technical books on plant taxonomy, and identify the appropriate botanical name. When possible, I also took the botanist along with me on treks; however, even then, he found the photographs useful for later verification.

It was necessary to make sure that the photographs would contain the details necessary for classifying plants. The main information botanists need is a view of the way in which different parts of the plants are arranged together, such as the way petals and sepals are grouped in a flower, or the way leaves grow on a stalk. They look also at the colour and shape of various parts of a plant, as well as subtle features such as whether the stem is ribbed, or has a smooth or hairy surface. Therefore, I took a set of photographs of each plant that aimed to include the views listed below:

1. an overall view of the plant showing it within its environment;
2. a whole stalk or branch, showing the arrangement of leaves, together with a close-up of a leaf, showing veins and other characteristics;
3. if flowers were present, an overall view of the flower, together with a close-up of the arrangement of sepals, petals, stamen and/or stigma, and other characteristics;
4. if fruit were present, an overall view of the fruit, together with the inside of the fruit (after cutting the fruit open);
5. any other plant characteristics.

Sample images are given in Figure 3–4. Figures 3a–c are of the pandlpëyriyë plant (*Lycopersicum esculentum*), showing views 1, 2, and 3, respectively. Figure 4 shows a striking characteristic of the biŋgi tree (*Pterocarpus marsupium*) – the red sap secreted when cut, which makes it appear as if the tree is bleeding.

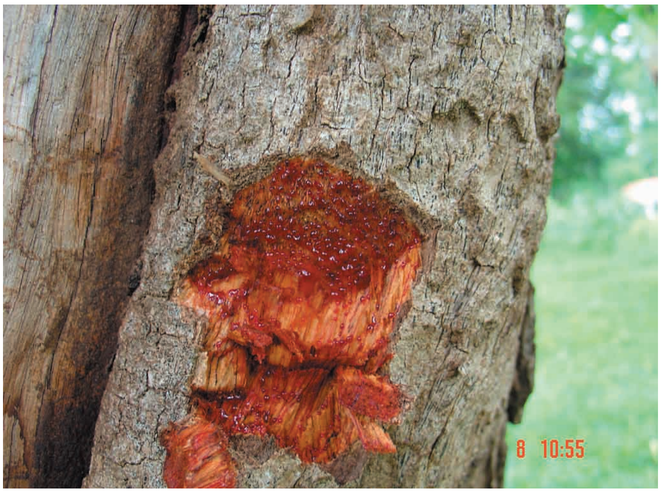
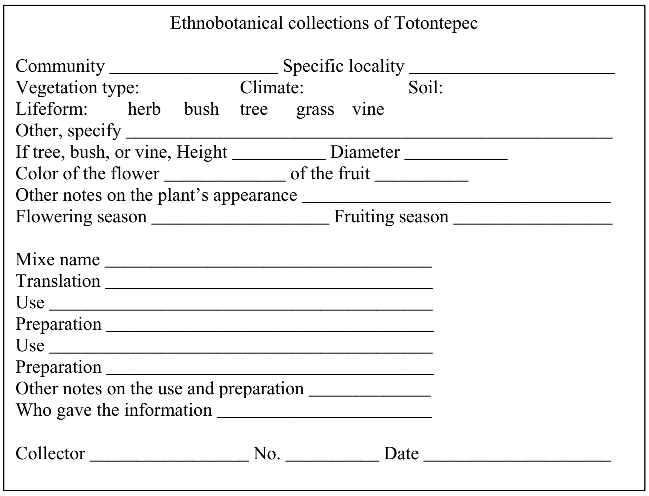


Figure 4: Blood-red sap secreted by the biŋgi tree (Pterocarpus marsupium) *(cropped by the editor)*



Ethnobotanical collections of Totontepec

Community ________________ Specific locality ____________________
Vegetation type: Climate: Soil:
Lifeform: herb bush tree grass vine
Other, specify ____
If tree, bush, or vine, Height _________ Diameter __________
Color of the flower ___________ of the fruit _________
Other notes on the plant's appearance _____________________________
Flowering season _________________ Fruiting season ______________

Mixe name _______________________________
Translation _______________________________
Use ____________________________________
Preparation _______________________________
Use ____________________________________
Preparation _______________________________
Other notes on the use and preparation ____________
Who gave the information ____________________

Collector _______________ No. _________ Date __________________

Figure 5: A sample from the pre-printed ethnobotanical notebook employed in Totontepec (Martin 2004:37)

Although this data collection method produced good results on the whole, it did leave a few issues unresolved. A useful improvement that I could have made would have been to place a ruler next to the relevant section of the plant, so that the photograph recorded its scale. Some plants, such as ferns and mushrooms, cannot be effectively identified from photographs and my botanical consultant advised that it would be necessary to have specimens studied in a laboratory. Since we had not asked relevant forest authorities for explicit permission to collect plant specimens, we had to leave this for future research.

In addition to photographs, it is advisable to note any information that consultants can provide about plants and animals. Martin (2004) suggests that ethnobotanists use forms of the type shown in Figure 5 (a copy of the form used for collecting Mixe plant names in Totontepec). He suggests assigning a unique number to each plant and filling in a separate form for each one; see his ethnobotany manual for further explanation. The form could easily be adapted for recording ethnozoological information.

Collecting animal names

To elicit animal names, I was limited to using books with good pictures, since it is difficult to spot animals in the countryside, much less photograph them. As my consultant would indicate an animal, I noted down both the name in the relevant language and the corresponding zoological name given in the book. This method worked better for animals that have clearly visible distinguishing features: large mammals like elephants, tigers, bears, or deer. Pictures of smaller animals such as birds, snakes, or fish present some of the same problems as those described for plants above because different species typically show, to the untrained eye, only slight differences in form or colour. Thus, consultants might point to a bird in the book and give me their name for it, while the book indicated that the bird was not found in the Nilgiri region. These mistakes did not arise, of course, from consultants' lack of knowledge about local bird species, but from the fact that books cannot present the object with its entire real-world context. My consultants rely not only on visual characteristics, but also on contextual cues such as the location where the animal is spotted, the way in which it moves, or the sound of its call.

I expect that the problems in asking consultants to name animals from books could be mitigated by asking a qualified zoologist to assist in the elicitation; however, time constraints prevented me from doing this. A zoologist is better equipped to ask consultants suitable questions about animal behavior, and to use the answers to correctly identify the animal. It is still, however, a good idea to take photographs where possible showing animals in their natural habitat or how the community utilises animal resources.

Equipment

I recommend the following equipment for collecting data such as discussed in this paper:

- Camera: a digital SLR camera with at least 5 megapixels resolution, with macro lens for photographing details and tiny objects, and telephoto lens for foliage of tall trees.
- Audio recording equipment: a quality digital recorder, tie-clip microphone, and radio-microphone kit. The latter allows the consultant to be mobile without getting cables caught in bushes (alternatively, the consultant can carry the recorder in a suitable bag).
- Video recording equipment: I did not attempt to videotape these treks; however, it would be good to videotape the consultants as they discuss the plants. Since the linguist will be absorbed in noting information, an additional person should be present to handle the video camera.

- Stationery: clipboard, notepages, data-entry forms, ruler (for measuring specimen sizes), etc.
- Carrying equipment: a backpack and waterproof vest or shoulder-bag with plenty of pockets to organise equipment and keep it safe from jolts and rain. Carry as much as possible in bags to keep the hands free for navigating safely through dense bushes.
- Safety and protective equipment: rainwear, clothing to protect against thorns, and shoes with thick leggings to protect against leeches. Silica gel crystals kept with equipment can protect it against moisture, but the packets should not leak crystals or powder onto the equipment.

Acknowledgements

Scientists at the Indian Institute of Science's field station at Masinagudy assisted me in identifying biological names for local flora and fauna. I am particularly grateful to R. P. Harisha and R. Arumugam. Harisha provided botanical names for the plants I photographed, accompanied me on some treks, and made it possible to visit the semievergreen forest plot shown in Figure 2. Arumugam provided the maps used. I am also very grateful to Masayoshi Shibatani, at Rice University, Texas, and to the Shell Center for Sustainability at Rice University for supporting this research.

References

Martin, G. (2004). Ethnobotany: A Methods Manual (People and plants conservation series). London: Earthscan.

Archiving

IMDI Metadata Field Usage at MPI

Alex Klassmann, Freddy Offenga,
Daan Broeder, Roman Skiba
MPI, Nijmegen

Introduction

Metadata is indispensable for discovering and searching the ever-growing volume of online language resources. Three metadata standards are now widely used for language resources – TEI, OLAC, and IMDI (links below). TEI is the oldest of these; OLAC was developed as an extension of the Dublin Core (DC) set which is widely used by librarians and for generalised cataloguing of web documents. The IMDI set was designed in collaboration with linguists, speech engineers and others to serve the specific needs of those researchers, especially resource discovery and retrieval, and is correspondingly more comprehensive.

An implicit purpose, therefore, of using IMDI is to support more accurate retrieval of resources. In reality, however, this can only be achieved if the metadata fields are actually accurately populated with searchable content. A large number of empty or inappropriately filled-in fields would prevent enhanced retrieval. Therefore, we saw that, after six years in operation, it would be very interesting to analyse our depositors' usage of IMDI. From such a study, we felt we could better understand:

- how well searches are working (searches that depend on poorly used elements may lead to wrong interpretations);
- where researchers find it difficult to enter descriptions and where, therefore, improvements to IMDI could be made;
- why some researchers complain about the necessity to create metadata (for example, some PhD students complain that time pressures do not allow them sufficient time to do so).

We focused on metadata descriptions that were created by individuals or small projects from the MPI and from DoBeS teams. Corpora where metadata was completely or partly generated, such as the Dutch Spoken Corpus, were excluded from the study. A total of 23,710 metadata description files were analysed.

Results

Figure 1 gives an overview of depositors' usage of IMDI fields (where usage means that some data has been filled in). A number of observations can be made:

- At the session level, project, geographic and date information is filled in for about 90% of cases, but descriptions with further useful information are provided in only 40% of cases.
- The description field at the content level is used in more than 70% of cases. At this level, depositors prefer to use this free-text description field rather than the Content Type fields such as Genre (30%), Sub-Genre (25%) or Subject (10%). The modalities in focus and the communicative context are used in more than 75% of cases.
- The language name is filled in in almost 100% of the cases. However, the language code was used in only 40% of cases, even though many of the codes can be selected from supplied lists.
- Information is frequently provided about actors. On average there are three actors (including the creator) per resource bundle. Language skills of informants are filled in in many cases, but information about sex, age etc. is very limited.
- Information about references, formats and types is available for almost 100% of resources. This means that the IMDI records do indeed act as a kind of glue bundling together files. In addition, we can use such information to automatically check consistency, e.g. for correct file extensions. Some fields such as file size are little used but could be filled in automatically.

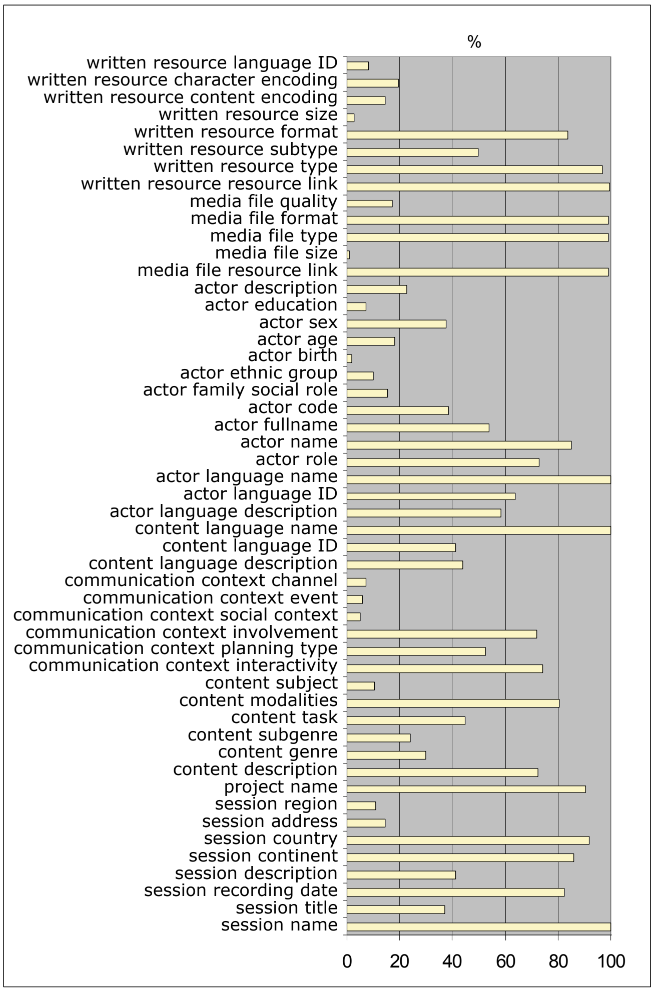


Figure 1. Usage of the most relevant IMDI fields, showing the proportion of IMDI files for which the field was filled in by depositors (from Klassman et al., 2006).

Discussion

The poor usage of the content type fields is somewhat disappointing. Local discussions have revealed that depositors find it difficult to use the built-in vocabularies and value sets, and that they have problems with categorising their resources. Some depositors did not know how to use these fields, or that it is possible to select multiple values. We doubt whether ad hoc changes in the value sets of the Genre, Subgenre or Subject fields will improve the situation, since we have to conclude that there is no commonly accepted vocabulary for them (except for some very basic terms). At a recent DoBeS workshop (June 2006), some researchers argued that classifications in endangered languages documentation would be more useful if they included genre vocabularies as understood by the language communities themselves.

The statistics also made us look in more detail at the use of the language name and language code fields. It was not clear to us why there was such a large discrepancy between the rate of usage of language names and language codes. Further investigation showed that in most cases it was possible to select a suitable language code. In fact, we developed scripts to correct obvious mistakes and add missing entries in these fields. As a result, the language name and code fields are now filled in and consistent in almost 100% of cases. We assume that depositors are either unfamiliar with the Ethnologue language codes or that they do not feel comfortable using them.

We concluded that the most important IMDI fields such as location, language, and recording date are used at a satisfactory level across all resources. For other fields, it seems that usage is dependent on the type of collection, and a more elaborate analysis needs to be carried out. However, the study has made clear that the description of the content type (Genre, Subject, etc.) can't be done at a satisfactory level. We were not yet able to draw any conclusions about particular IMDI fields that might reasonably be eliminated.

References and links

Klassmann, A., Offenga, F., Broeder, D., Skiba, R., Wittenburg, P. (2006). Comparison of Resource Discovery Methods. In: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik & D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 113–116. Paris: European Language Resource Association.

Wittenburg, P., Peters, W., Broeder, D. (2002). Metadata Proposals for Corpora and Lexica. In: M. G. Rodriguez & C. P. S. Araujo (Eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1321–1326. Paris: European Language Resource Association.

TEI: http://www.tei-c.org/

OLAC: http://www.language-archives.org/OLAC/metadata.html

Dublin Core http://dublincore.org/

IMDI: http://www.mpi.nl/IMDI

Unique Resource Identifiers

Daan Broeder, Eric Auer, Peter Wittenburg
MPI, Nijmegen

Many valuable language resources have become available via the Internet. These include primary resources such as media recordings, secondary resources such as annotations, lexica and grammars, and other electronic publications. Increasingly, there are also references and links between resources: entries in lexica refer to fragments in annotated media recordings, and grammars refer to examples in primary sources.

Currently, the typical way of making such links is via hyperlinks based on URLs (Uniform Resource Locators). However, URLs are unstable identifiers, compared, for example, to bibliographic references used for books and journals. URLs can change as servers and storage systems are replaced or reorganised by system managers, or as resources are moved between institutions. And when such changes do occur, it is practically impossible to modify all existing references to URLs. Therefore, URLs do not provide a stable and persistent method of referencing.

Unique Resource Identifiers (URIDs)

We need a mechanism that can overcome this instability. Valid references need to be maintained while the locations of the resources change. A solution is to introduce a global naming system that uniquely and persistently identifies each electronic resource separately from its URL address. We call the identifiers "Unique Resource Identifiers" or URIDs. These can be compared to ISBN numbers; they uniquely identify titles and provide a distinction between the information object and its physical instances at particular addresses.

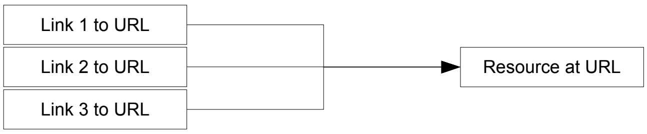


Figure 1. Links referring to physical addresses

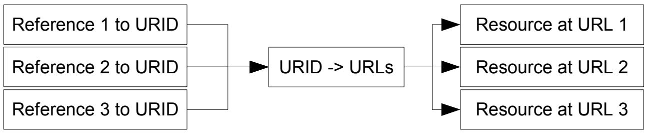


Figure 2. Links referring to URIDs

Like ISBNs, URIDs are administered by institutional registries. These registries provide the service of translating URIDs into actual web addresses, so that the resources can be accessed. This mechanism can again be compared to the world of books: one looks up a book's title but then has to get an "address" (i.e. the book's library call number) in order to locate it on the shelves before reading it. In other words, using URIDs means introducing the complication of another layer of reference between resource identities and their instances. However, the advantage is that when resources move, the addresses only have to be administered in one location – the URID registry.

In the existing system of web links, several links may point to a given resource at its physical address (Figure 1). By contrast, in a URID system, all links refer to a unique and persistent identifier, which in turn is mapped to one or more URLs (Figure 2). Incidentally, the figure illustrates the case where there are multiple instances of the same information resource, for example in different archives.

The main challenges in introducing this additional layer into the linking mechanism are that the mappings of URIDs to URLs have to be done with care, and that the online services that provide the mappings have to be available 100% of the time.

DAM-LR and the Handle system

In the DAM-LR project (Distributed Access Management for Language Resources), the participant archives chose an architecture that combines the advantages of URIDs with the flexibility to cater for the particularities of each archive. We chose the Handle System, already widely accepted in the world of digital libraries and archives, as our URID (or "handle") registry and resolving (mapping) system. A complete URID (handle) is a combination of a centrally-assigned prefix and a locally-defined postfix. The Handle System registers institutions as local authorities with their own URID prefix. The institutions can then locally specify the remainder of the identifiers (the "postfixes"), map the identifiers to URL paths, and provide a public service mapping handles to paths. The global Handle System authority directs references that cross prefix domains to the appropriate local mapping service.

Next steps

Participants in the DAM-LR project registered as authorities at the Global Handle Registry service. This allows them to maintain their own mapping services and operate independently, while still enjoying the benefits of URIDs.

URIDs as described above refer to complete resources. We would ultimately also like to be able to refer to relevant fragments within resources, such as segments of video or entries in a lexicon. For XML-encoded text documents, handles could be combined with XPointer references to achieve this; however, a general mechanism is yet to be defined. Nevertheless, URIDs provide the first step in solving the question of how to create a reliable system for referring to electronic documents.

Links

DAM-LR: http://www.mpi.nl/dam-lr

Handle: http://www.handle.net/

News in Brief

Language Archives Inspire Publishing Donation

Pauline Edwards
FPCF, Brentwood Bay

FirstVoices.com is a set of web-based language archives and teaching resources, developed by First Peoples' Cultural Foundation – a Canadian-based Indigenous non-profit society, based in British Columbia. At FirstVoices, each language community administers their own archives by uploading word and phrase lists, songs, stories, pictures, audio files and videos.

FirstVoices is managed by a dedicated team, assisting Indigenous communities in the archiving process with training, technical advice and ongoing support to promote and revitalize language learning. Custom software keyboard solutions for virtually any Indigenous language are being developed and distributed by FirstVoices, enabling Indigenous communities to easily communicate using their own language.

Recent exposure for FirstVoices.com at international conferences in Canada, Japan and Botswana are raising the profile of the unique language tools, originally developed for the 198 First Nations in BC. A recent invitation to showcase FirstVoices.com at the second World Information Technology Forum (WITFOR) in Gaborone, Botswana, acknowledges the successful development and implementation of a made-in-Canada technology solution developed by Indigenous people, for Indigenous people.

To advance this important global cause, Trafford Publishing is making a donation of approximately $1.6 million in publishing costs over the next ten years. Authors of Indigenous language books will receive Trafford's "Best Seller Plus" package, worth $2549. To request an information kit and application form, Indigenous language teams are invited to contact the Trafford FirstVoices Publishing Program coordinator Pauline Edwards. Reach her at pauline@fpcf.ca.

Links

First Voices: http://www.firstvoices.com

Trafford Publishing: http://www.trafford.com

DAM-LR Training and Workshop Lund, Jan 2006

Sven Strömqvist
Centre for languages and literature, Lund University

During January 26–28 this year, the project Distributed Archive Management of Language Resources (DAM-LR) ran a combined workshop and training event at Lund University. The first part of the workshop addressed local architectures and archive solutions at Max Planck Institute for Psycholinguistics, Nijmegen; School of Oriental and African Studies, London; Instituut voor Nederlandse Lexicologie, Leiden; and the Centre for languages and literature, Lund. Special attention was given to metadata specifications (OLAC, IMDI), and the structure of the federation of participating institutes and archives. A full-fledged server for the DAM-LR project was launched, with metadata and an authentication and authorisation system.

The second part of the event consisted of a training session for DAM-LR personnel together with participants from other institutions, focusing on archive building, IMDI metadata specifications, and the DAM-LR access system. Altogether, some 30 experts, ranging from researchers to university librarians and IT engineers, participated in the session.

The workshop has had several positive effects. First, the training event inspired several researchers to take measures to make their research data publicly accessible. Thus, a large set of audio and video recordings, photographs, field notes and drawings of Kammu (Laos) language and culture over the past 30 years – which had been previously dispersed in shelves and drawers – are now being digitized, classified using IMDI, and made publicly accessible. Second, the Lund University library has signalled an increased interest in cooperating with DAM-LR to extend their high-end electronic information services with research data from the language sciences.

MPI Archive Services Expanded to New Audiences

Paul Trilsbeek, Jacquelijn Ringersma, Peter Wittenburg
MPI, Nijmegen

MPI archivists offer three services to interested researchers, projects and institutions:

- storing resources in the MPI archive
- setting up local archive systems
- carrying out local training courses

Storing resources

Many digital (and non-digital) language resources are at risk of loss or degradation due to inappropriate storage. Deposit with specialised archives such as at MPI significantly increases the chance of long-term survival of resources, and, equally importantly, makes them discoverable for others.

Researchers who have valuable language resources can now contact the MPI archivists about depositing their resources. Although the MPI expects to receive digital files for depositing, in certain cases the MPI might also be able to assist with digitisation/capturing. Together with the depositor, the MPI will check the state of the materials and how much time will be needed to ingest them into the archive. It is expected that depositors organise the structure of their resources and participate in the creation of metadata descriptions for them. It is also expected that most materials will have the depositor's permission to be made available to others for research purposes; however, depositors will retain control over access-granting procedures for their materials.

This policy is also designed to encourage researchers working on MPI-related projects (e.g. DOBES projects) to continue archiving new materials after their project has ended.

Setting up local archives

We can set up local archives using MPI-developed software. Primarily this includes LAMUS (Language Archive Management and Upload System), AMS (Access Management System), and the IMDI (ISLE Metadata Initiative) infrastructure. This setup allows archivists to ingest new resources into their archive, to manage access policies, and to provide user access to resources via the web. The setup can be extended (if desired) by installing content access applications such as ANNEX (access to annotated media streams) and LEXUS (access to multimedia lexica). Experts from the MPI set up the software, provide training for archivists and system managers, and finally hand over management of the system. If desired, we can set up a dynamic link to ensure that local changes are also applied to a mirror site at the MPI, thereby fully integrating the local archive into MPI's long-term preservation strategy.

Training courses

MPI archivists can provide local training courses on various technical topics related to fieldwork and archiving. Courses, at various levels, are based on DoBeS training courses; see http://www.mpi.nl/DOBES/training_courses/.

Limitations

The number of services that we are able to provide per year is limited. Requests are dealt with on a first come, first served basis.

Announcements

Synpathy Syntax Editor

The MPI for Psycholinguistics has developed a graphical syntax editor called Synpathy, based on the SyntaxViewer from the TIGERSearch project. It allows users to select a sentence from a corpus, graphically design a syntax tree, and to create an XML output in theTIGER format. These descriptions can be visualised within ELAN and ANNEX in synchronisation with their media files. A manual is currently being created. Synpathy is available as a stand-alone Java tool.

Links

Synpathy:
http://www.mpi.nl/tools/synpathy.html

TIGERSearch:
http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

DELAMAN IV at ELAR

The Fourth Annual meeting of DELAMAN (Digital Endangered Languages Archives and Musics Archive Network) is being held at ELAR, SOAS in London, 2–3 November 2006. Participants and topics are listed on the website below. Topics of particular interest to linguists and digital archives include:

- Towards achieving academic recognition of archived language resources – developing and disseminating citation conventions
- Media overload: the high price of digital media for language documentation archives. How do we measure value?

Reports on these topics will appear in the next issue of LAN. For more details on the meeting, see http://www.delaman.org/meeting2006.html

LANGUAGE
ARCHIVES
NEWSLETTER

Web: http://www.mpi.nl/LAN/
Mail: LAN@mpi.nl
ISSN: 1573-4315

Editors: David Nathan (ELAR, SOAS, London)
Paul Trilsbeek (MPI, Nijmegen)
Marcus Uneson (Lund University)
Layout: Marcus Uneson

Contributions welcomed at: **LAN@mpi.nl**

Last submission date for the next issue: **November 25, 2006**