# The Language Archive – a new hub for language resources

**Sebastian Drude, Daan Broeder, Paul Trilsbeek, Peter Wittenburg**
Max Planck Institute for Psycholinguistics
P.O. Box 310 · 6500 AH Nijmegen · The Netherlands
E-mail: {Sebastian.Drude, Daan.Broeder, Paul.Trilsbeek, Peter.Wittenburg}@mpi.nl

### Abstract

This contribution presents "The Language Archive" (TLA), a new unit at the MPI for Psycholinguistics, discussing the current developments in management of scientific data, considering the need for new data research infrastructures.
Although several initiatives worldwide in the realm of language resources aim at the integration, preservation and mobilization of research data, the state of such scientific data is still often problematic. Data are often not well organized and archived and not described by metadata – even unique data such as field-work observational data on endangered languages is still mostly on perishable carriers. New data centres are needed that provide trusted, quality-reviewed, persistent services and suitable tools and that take legal and ethical issues seriously. The CLARIN initiative has established criteria for suitable centres.
TLA is in a good position to be one of such centres. It is based on three essential pillars: (1) A data archive; (2) management, access and annotation tools; (3) archiving and software expertise for collaborative projects. The archive hosts mostly observational data on small languages worldwide and language acquisition data, but also data resulting from experiments.

**Keywords:** data centre; archiving; research infrastructure

## 1. State of data and challenges

Worldwide, there are currently big efforts going on to make data more visible, accessible and interoperable. For instance, the European Strategy Forum on Research Infrastructures (ESFRI) process has led to over 40 initiatives (ESFRI Initatives, all underlined terms refer to entries in the references), the European Commission (EC) is funding domain specific and horizontally organized data infrastructures (EC Data Infrastructures), the American National Science Foundation has launched the DataNet initiative, and the Australian government has set up the Australian National Data Service (ANDS), all with similar goals. Hence, also research and funding organizations are increasingly aware of the need of proper data management, curation and preservation, worldwide and across all disciplines.

In the area of the humanities, projects and institutions, with different degrees of geographical and thematic coverage, such as CLARIN (one of the ESFRI Initiatives), MetaNet, FlareNet, ISO TC37/SC4, DOBES, Paradisec, AILLA and many more language resources (LR) related initiatives want, amongst other goals, to tackle the mentioned challenges.

Still, currently the state of scientific data is often problematic. Projects still keep data on non-accessible containers and they are frequently not well organized. Backups are made in an improvised and ad-hoc manner, and the need for sustainable access to research data, including changing data formats, is only seldom addressed. In the on-going dramatic turn towards data orientation in all scientific domains, this situation is recognized as alarming and the pressure on all stakeholders to change it is increasing. However, several aspects hamper progress: (1) A change of culture in researchers' mind is demanding; new levels of trust need to be established and career building incentives created. (2) We lack stable and persistent data hubs offering reliable and trustful services, easy to integrate into the researchers' daily workflow. (3) We lack tools and standards that are transparent to the researchers, but nevertheless effectively and efficiently support the above mentioned goals. (4) We lack a new type of "data scientists" that understand the digital era and its mechanisms, acting as data managers and curators. The report for the EC "Riding the Wave" (High Level Expert Group, 2010) summarizes the important points.

An example from the LR domain may illustrate this state of affairs. Currently we have about 6500 languages spoken worldwide. Estimates are that this figure may shrink to 600–1200 in the next 140 years (Kraus, 1992). That is, at least every two weeks one language disappears, and with each of them an enormous treasure of knowledge about linguistic systems, cultures and aspects relevant for health and nature. Not only have we the obligation to document the current diversity and preserve this heritage for future generations, but this documented material should also be available for revitalization programs, and for research. For example, to understand our mind's language faculty we must consider the huge linguistic diversity worldwide. Language systems can differ so much that Levinson and Evans (2009) doubt whether language universals exist at all.

However, a UNESCO study (Schüller, 2004) shows that about 80 % of the ethno-linguistic recordings are highly endangered due to material deterioration and media stored in researchers' cupboards, and, more dramatically, due to fragile formats and inappropriate technology.

## 2. Data Centres and Requirements

Setting up strong persistent data centres (DCs) is of utmost importance in overcoming this situation in all sciences and the humanities. Of course such DCs must fulfil a number of requirements to be successful: (1) Being service oriented and not having an own agenda for the data: this would destroy trust. (2) Being stable and persistent: without a guarantee to be able to access the data at all moments for many years, researchers may not take the effort of transferring the data. (3) Having "data scientists" that know about standards and methods and that can take essential curation steps to maintain easy accessibility. (4) Offering tools allowing researchers to embed the stored data in their workflows and, increasingly often, chains of automatic operations. (5) Taking care of legal and ethical rules which often appear obscure in the worldwide data networks.

In addition, there are formal requirements data centres must meet. For instance, CLARIN established criteria for its data centres (CLARIN requirements) which apparently are widely accepted. In CLARIN, data centres need to:

- offer useful services to the community and to agree with basic principles: explicit statements about the duration and the quality of the services, guarantee to deliver the same content for the same identifier, adherence to standard interfaces, protocols and agreements to achieve a high degree of integration and interoperability
- adhere to security guidelines, i.e. their servers / services have accepted certificates, and participate in data federations allowing users to create virtual collections
- have a proper, clearly specified repository system and participate in a quality audit procedure
- associate persistent identifiers with their resources that can be used to proof integrity and authenticity
- offer appropriate metadata (via OAI PMH)

The CLARIN initiative revealed that it is not generally easy for data holders to meet these criteria, in particular due to lack of appropriate funding and expertise. Pressure from research organizations and funders, however, will grow, and some of the centres may fail while others will remain as strong data centres certified according to Repository Audit and Certification (RAC) (Conrad 2010) or the Data Seal of Approval (DSA). Perhaps the latter will offer data collected from various disciplines or just data from a certain domain. Here factors such as economy of scale can be applied for bit-stream preservation, but hardly for data curation. It is even in debate whether commercial services can or should be used to establish such centres. Currently, however, most of the European researchers are afraid that commercial centres will eventually give access only to those having sufficient funds, whereas publically funded centres should offer data for free to all researchers, including to the so-called data scientists. However, with the current mostly project-based funding schemas, the challenges to maintain data centres and networks in a long perspective are considerable.

In the area of LR, with the emerging e-Research paradigm, also existing dedicated centres focusing on academic research – such as the Evaluations and Language Resources Distribution Agency (ELDA), the Linguistic Data Consortium (LDC), the Institut für Deutsche Sprache (IDS) or the Instituut voor Nederlandse Lexicologie (INL), among many others – may have to adapt their procedures and policies in view of the need for easy, transparent, selective and interactive use of language resources, and to enhance their strategies with respect to long-term preservation.

## 3. Conception and goals of The Language Archive

In view of this scenario and developments, the Max Planck Society (MPG), the Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Sciences (KNAW) joined forces and established a new unit at the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands (MPI-PL): "The Language Archive" (TLA). The new unit continues and consolidates the well-known work by the Technical Group (TG) at the MPI-PL under its technical director Peter Wittenburg. The TG functioned already as archive and technical centre of the DOBES program and had an important role in several other of the above-mentioned developments and initiatives, in particular in CLARIN.

A clear signal was to be given that TLA will become a strong, persistent and trustful data centre, offering its collaboration to any researcher with serious language data of all sorts. The funding from the three institutions supports six permanent positions at TLA as a basis for this new unit. Still, the lion's share (currently about 20 technologists) is financed by third and project money.

TLA is based on three essential pillars: (1) A data archive holding resources on languages and cultures worldwide. (2) Management and access tools developed and maintained in collaboration with a wide variety of projects. (3) Archiving and software expertise for collaborative projects.

The primary goal of TLA is to store and preserve digital language resources, to give access to researchers and other interested users and to develop and integrate new technologies advancing language research. Although TLA will be primarily grounded on the research needs of the MPG, BBAW and KNAW, it is open to researchers and "citizen scientists", open to all requests for depositing any suitable language related data. The current focus is on observational data from languages all around the world (typically manually annotated audio and video recordings), such as data from the DOBES program or other data resulting from ethno-linguistic field research, but also observational data from language acquisition studies, mainly for major better studied languages.

TLA also hosts data resulting from experiments: psycholinguistic studies with response patterns to stimuli, eye-tracking data, and more recently also neurological imaging data connected with language production or perception, and even genetic data related to linguistic topics. In principle, any well-structured digital data (no physical objects, data carriers are returned to the depositors) with long-term linguistic scientific relevance are accepted.

TLA promotes a culture of free sharing of data and believes that in principal and wherever possible data should be made freely accessible via central online repositories. However, the personal and privacy rights of the speakers (and sometimes the intellectual property rights of the depositors) have to be taken into account so that controlled or regulated access may be necessary for certain parts and/or types of data (for instance sensitive material such as sacred rituals). Therefore efficient mechanisms have been created that allow depositors and managers to easily associate rights with individual data objects or branches in a metadata tree, possibly restricted to certain data types (audio, video, annotation etc.). Four levels of access have been established: (1) open, (2) open to registered users, (3) request based, (4) closed. Except for the first, all may require signing a Code of Conduct (such as in the case of DOBES data). Currently, about 25% of the resources are of level (1) or (2). Access to most other material can be requested. These protection mechanisms are seen as sufficient – for various reasons, logo introduction, watermarking, encryption etc. are not applied.

The tools developed at TLA so far focus mainly on two areas: On the one hand, on the needs for archiving such language data – the by now well-known web-oriented tools LAMUS, AMS, ARBIL, IMDI-browser, TROVA and ANNEX will be maintained, further developed and updated. For instance, the technology based on the IMDI-metadata-standard used so far for all data archived at TLA will be substituted by newer versions based on the new component-meta-data-infrastructure (CMDI) being developed in the CLARIN initiative. On the other hand, tools needed for linguists concerned with the linguistic diversity worldwide by documenting and describing individual languages, usual in a field research setting. In this context, the now widely used ELAN annotation software has been developed as well as the LEXUS online tool for the creation of multi-media lexical databases. Also these tools will continue to be maintained, developed and enhanced, for instance by developing or integrating new computational methodologies for the study of languages such as new types of audio/video pattern recognizers making multimodal work more efficient, or analysis tools that extract regularity features from given annotations.

Participating in national and international projects and collaborations, TLA contributes to the emerging e Research infrastructures, advancing and promoting international standards. TLA will participate in initiatives developing and maintaining advanced software to allow archive managers to organize and maintain a consistent and coherent digital archive. State of the art software (see above) will also allow users to easily create, access, and enrich the data stored at TLA. Via CLARIN and similar initiatives, TLA cooperates with other data centres and promotes the integration of language resources worldwide, but besides this, no specific agreements with individual initiatives or institutions such as LDC, FlareNet or MetaNet have been made so far.

TLA is devoted to fulfil the mentioned requirements established by CLARIN. In particular, TLA will continue to assess its archival practices by applying the DSA process regularly. It will also participate in European data infrastructures to promote cross-disciplinary data access and in the worldwide Data Access and Interoperability Task Force (DAITF).

## 4. Conclusions

With the foundation of TLA the three participating institutions have set a signal to support the current trends towards strong, reliable and persistent non-profit data centres. The new unit is based on what has already been done by the Technical Group at the MPI-PL for more than two decades and will intensify this work, based on increased structural funds. For the essential task to establish proper trust relations with all depositors and towards the community, it will respect the copyright of the data depositors, and take all required steps to curate and preserve the stored data.

The establishment of TLA will increase the capability of ensuring long-term support for the tools being developed. This incentivizes users to work with a complex tool set. Perhaps TLA will be able to take over software from others to improve its availability. We expect that TLA can contribute to making language resources more visible and accessible and to better preserve them over time.

## 5. References

AILLA: The Archive of Indigenous Languages of Latin America. www.ailla.utexas.org. Last visited 23.3.2012.

ANDS: Australian National Data Service. www.ands.org.au/. Last visited 23.3.2012.

BBAW: Berlin-Brandenburgische Akademie der Wissenschaften. www.bbaw.de. Last visited 23.3.2012.

CLARIN: Common Language Resources and Technology Infrastructure. www.clarin.eu. Last visited 23.3.2012.

CLARIN Requirements: CLARIN (2010). Criteria for CLARIN Centres. Online at: http://www.clarin.eu/system/files/centre_criteria-CLARIN-ShortGuide.pdf. Last visited 23.3.2012.

CMDI: Component MetaData Infrastructure. www.clarin.eu/cmdi. Last visited 23.3.2012.

Conrad, M. (2010). Note on ISO 16363. (RAC). Online at: mailman.ccsds.org/pipermail/moims-rac/2010-May/000826.html. Last visited 23.3.2012. RAC Wiki: wiki.digitalrepositoryauditandcertification.org. Last visited 23.3.2012.

DAITF: Data Access and Interoperability Taskforce. www.daitf.org. Last visited 23.3.2012.

Datanet: Sustainable Digital Data Preservation and Access Network Partners (DataNet). Online at: www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141. Last visited 23.3.2012.

DOBES: Dokumentation Bedrohter Sprachen. www.mpi.nl/dobes. Last visited 23.3.2012.

DSA: Data Seal of Approval. www.datasealofapproval.org/. Last visited 23.3.2012.

EC Data Infrastructures: cordis.europa.eu/fetch?CALLER=PROJ_ICT&QZ_WEBSRCH=data+infrastructure&QM_PJA=&QM_PGA=FP7%24&USR_SORT=EP_PJA_A+CHAR+ASC. Last visited 18.2.2012.

ELDA: Evaluations and Language Resources Distribution Agency. http://www.elda.org/. Last visited 18.2.2012.

ESFRI Initatives: “European Strategy Forum on Research

Infrastructures" Network of RIs funded by the European Commission under FP7. Online at: ec.europa.eu/research/infrastructures/index_en.cfm?pg=ri_projects_fp7. Last visited 23.3.2012.

FLareNet: Fostering Language Resources Network. www.flarenet.eu. Last visited 23.3.2012.

High Level Expert Group on Scientific Data (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data*. A submission to the European Commission. Online at: cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf. Last visited 23.3.2012.

IDS: Institut für Deutsche Sprache. www.ids-mannheim.de/. Last visited 23.3.2012.

INL: Instituut voor Nederlandse Lexicologie. www.inl.nl/. Last visited 23.3.2012.

ISO TC37/SC4: International Organization for Standardization, Technical Committee 37, Subcommittee 4, "Terminology and other language and content resources". www.tc37sc4.org/. Last visited 23.3.2012.

KNAW: Koninklijke Nederlandse Akademie van Wetenschappen. www.knaw.nl. Last visited 23.3.2012.

Krauss, M. (1992). The World's Languages in Crisis. *Language* 68, pp. 4–10.

LDC: Linguistic Data Consortium. www.ldc.upenn.edu/. Last visited 23.3.2012.

Levinson, S. and Evans, N. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32, pp. 429–492. doi:10.1017/S0140525X0999094X.

MetaNet: Multilingual Europe Technology Alliance Network. www.meta-net.eu. Last visited 23.3.2012.

MPG: Max-Planck-Gesellschaft. www.mpg.de. Last visited 23.3.2012.

MPI-PL: Max Planck Institute for Psycholinguistics. http://www.mpi.nl/. Last visited 23.3.2012.

OAI-PMH: Open Archives Initiative (w/o. d.) Protocol for Metadata Harvesting. Online at: http://www.openarchives.org/OAI/openarchivesprotocol.html. Last visited 23.3.2012.

Paradisec: Pacific and Regional Archive for Digital Sources in Endangered Cultures. paradisec.org.au. Last visited 23.3.2012.

Schüller, D. (2004). Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity. *Language Archive Newsletter*, 1(3), pp. 9–10. Online at: www.mpi.nl/lan/issues/lan_03.pdf. Last visited 23.3.2012.

TLA: The Language Archive at the Max-Planck Institute for Psycholinguistics. tla.mpi.nl. Last visited 23.3.2012.