Digital curation: the SLDR experience

Frédérique Bénard , Bernard Bel

Laboratoire Parole et Langage (LPL)
CNRS – Aix-Marseille University
B9 80975, 5 avenue Pasteur
13604 Aix-en-Provence, France
frederique.benard@lpl-aix.fr, bernard.bel@lpl-aix.fr

Abstract

This paper deals with the description, packaging, and preservation of digital objects submitted to the Speech & Language Data Repository (www.sldr.org). SLDR is a Trusted Digital Repository offering the sharing oral/linguistic data and its submission for medium-term and long-term preservation via an institutional archive. Its work environment offers a flexible integrated management of access rights at all phases of a project. Data include all signals associated with oral production, documents created or collected during an experiment or a field enquiry, material derived from primary data with their associated resources and tools designed for data processing in the domain. Currently, information packages are distributed via the Adonis/Huma-Num grid hosted by *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3) and preserved on the platform of *Centre informatique de l'enseignement supérieur* (CINES), a site beneficiary of the Data Seal of Approval.

Index Terms: digital curation, data repository, resource sharing, OAIS, archive, Digital Humanities

1. A change of practice with respect to archives

In recent years, funding agencies supportive of linguistic research projects have been putting pressure on scholars to include long-term preservation and sharing of data in their project agenda. Initiating the archiving process at the very onset of the project is a radical change of practice because of associated technical and legal constraints. It is made possible by tools and procedures compliant with the life-cycle of present-day research projects. These will be introduced in this paper.

Compliance of procedures is achieved by responding favourably to the requests of data producers. Among these, we give brief answers to the most challenging ones:

- *Should I wait for the completion of my research work to submit only final versions of data and results?* Answer: A digital repository offers the options of upgrading stored material either via the submission of new versions or correcting mutable data.
- *Should I wait for the availability of informed consent by all participants to share recording in open access?* Answer: Open access is one among many options. Access to sensitive data may be restricted (in compliance with the legal framework) and an integrated management of access rights facilitates their gradual modification in more or less restrictive directions.

2. A repository for sharing oral/linguistic resources

In 2006, the LPL laboratory was commissioned by the Social Science and Humanities department of the French *Centre national de la recherche scientifique* (CNRS, www.cnrs.fr) to set up a resource centre for speech research.

This initiative was driven by a growing concern with the existence of scattered oral resources in non-persistent formats and locations, many of which could not be reused nor shared due to access restrictions.

Another incentive was to promote the self-archiving of linguistic resources in a manner similar to that of scientific publications submitted to *Centre pour la communication scientifique directe* (CCSD, www.ccsd.cnrs.fr). In those days, the dissemination of speech corpora was mostly carried out by corporate agencies (ELDA, www.elda.org, and the LDC, www.ldc.upenn.edu). Admittedly, CNRS' initiative could initially be perceived as creating a public facility to replace a business framework unfit for academic work in small research units. Later on, this feeling of competition vanished thanks to a fair combination of private and public models (see infra §5.3).

At LPL, a generic (multidisciplinary and multilingual) digital repository was implemented from scratch after comparing existing similar initiatives [1]. During the same period, CNRS supported the creation of another site mostly replicating the design of LACITO's archive of rare languages [5].

In 2008-2011, LPL and LACITO were enrolled in a pilot project coordinated by TGE Adonis (www.tge-adonis.fr) for digital resource pooling in social sciences and the humanities. In this context, both repositories became submission sites for long-term preservation. After the completion of this project, LPL's resource centre was renamed *Speech & Language Data Repository* (SLDR) and LACITO's *COllections de COrpus Oraux Numériques* (COCOON).

A new phase of development started at the end of 2012. Six leading institutions joined efforts in the ORTOLANG project (www.ortolang.fr/english) with the aim of building a French sub-network of CLARIN centres (www.clarin.eu) involving SLDR for speech and CNRTL (www.cnrtl.fr) for text linguistics. Current focus is on interoperability with respect to descriptive metadata, persistent identifiers and controlled vocabularies.

3. From secure backup to long-term preservation

When initiating a research project, scholars should be aware of archival limitations with respect to file formats, and proceed to a comprehensive separation of primary and secondary data

with respective cycles of versioning. It is also advisable that files contained in a given folder share identical access restrictions.

These constraints are easier to comply with when secure backups are replaced with a process carrying out a technical validation of information package content and a set-up of controlled access to documents. In other words, medium-term preservation should follow the same procedures as long-term preservation.

As claimed on the CINES website (www.cines.fr), preserving digital resources is neither a backup service nor 'the ultimate step of storing data before oblivion or permanent loss.' In a long-term preservation scheme, data should be eligible for reuse after an unspecified period of time, typically more than 30 years. This calls for reliance on an institutional archive (CINES) rather than a consortium of computing centres whose policy might vary (because of fund scarcity) at a time data producers are no longer able to negotiate an extension of its preservation.

The commitment of CINES is threefold: (1) preserving data and its associated metadata; (2) preserving access-right information; (3) ensuring the reusability of data, which is achieved by migrating file formats (without loss of information) once these are becoming obsolete.

Digital items stored at the CINES archive are processed as generic 'information packages' regardless of their origin. Subject-specific information is stored in descriptive metadata encapsulated in XML files. Nonetheless, this implies technical limitations with respect to the packaging of Submission Information Packages (SIP) and a restricted set of open formats eligible for long-term preservation and logical data migration (www.sldr.org/wiki/Formats).

The issue of file formats is problematic when it comes to sound/video material. CINES accepts sound recordings in WAVE and AIFF with PCM encoding, or in compressed AAC — all non-proprietary formats. The popular MP3 format is not suitable because of its commercial restriction. Thus, if a corpus contains MP3 files, these will stored in medium-term preservation whereas their replications in WAVE or AIFF format can be submitted for long-term preservation.

Preserving video recordings requires a trade-off between accuracy (e.g. high-resolution, 3D or multiple cameras), reasonable storage space and eligible formats (currently MP4/AVC/AAC, OGG/Theora/Vorbis and MKV/AVC/Flac). Even though SLDR may accept items featuring thousands of files and sizes beyond 100 Gigabytes, it should be kept in mind that the current annual cost of long-term preservation is roughly 5,000 euros per Terabyte (CINES estimation).

4. Digital curation

The term 'digital curation' is a combination of 'digital preservation' and 'curation', the latter in the sense of 'activities that add value and knowledge to the collections' (Tammaro & Madrid, cited in [6, p. 2]). Combining these words reflects an evolution of archival practice made possible by the use of digitized documents as research material eligible for long-term preservation, dissemination and reuse outside their original production environment. This creates new opportunities for enriching data provided that the issue of its portability has been properly addressed [2].

As put by Hedstrom [4, p. 2], digital curation calls for expertise (and training) 'across a spectrum from curation-centric needs to discipline or application specific requirements'. The challenge is to fill the gap of expertise between producers (scholars, participants, informants) and archive curators in charge of the preservation of research/documentation material. On the side of archive curators, records management is no longer restricted to the preservation of 'semi-current' or 'inactive' records. It now includes 'curation at the source': assistance with the elaboration of descriptive metadata even before starting the collection of primary data.

Digital curation requires collaborative work and a proper coordination of initiatives to enrich archived material. To this effect, curation tasks accomplished by data depositors or administrators are traced by SLDR and displayed on the curation page (www.sldr.org/curation).

4.1. Packaging research data

The Huma-Num framework for long-term preservation of digital resources is based on the Open Archival Information System (OAIS) [3]. It comprises the two submission sites (SLDR and COCOON) connected with CINES (www.cines.fr) for long-term preservation and CC-IN2P3 (cc.in2p3.fr) for data dissemination, as shown Figure 1.

Items stored at SLDR are generic: any tree-structure of computer files with no limitation on size or (UTF8-encoded) file names. File formats incompliant with long-term preservation are automatically redirected to the dissemination site; this is the case for instance of sound files in MP3 format or video files in FLV format used for sound/video streaming, or ZIP files making it easy to download subsets of the tree.

SLDR deposit policy is the same as CCSD with respect to scientific papers (http://hal.archives-ouvertes.fr): documents, resources, corpora are not reviewed in terms of their presumed scientific or cultural-heritage value. The same permissiveness applies to the assessment of acoustic quality, accuracy of transcriptions, relevance of annotations etc., all of which should be taken care of by data producers. For this reason we incite producers to specify institutions responsible for data production and verification, as well as the funding bodies associated with the project (see Figure 2). Digital curation only cares for the proper technical packaging of data.

In a near future, ORTOLANG will develop facilities for the analysis and proper reuse of archived material. At this stage, the issue of data quality will be raised and guidelines produced to optimize interoperability and facilitate automated linguistic analytical processes.

4.2. Descriptive metadata

Every item has a specific space for 'documentary files' which may include documents describing its content, experimental protocols, associated material (diagrams etc.), transcriptions, translations and annotations of sound/video recordings etc. If the item is stored in medium-term or long-term preservation, documentary files will be modified without resubmitting a new version of the same item.

Documentary files include descriptive metadata, i.e. structured sets of information describing the contents of an item and its associated documents. Metadata are stored as XML files (at least one for each item) in the archive. In addition, the same are available in a database for quick access and reformatting.

4.2.1. Dublin Core OLAC

This metadata format is a qualification of Dublin Core applicable to the description of linguistic resources. See for

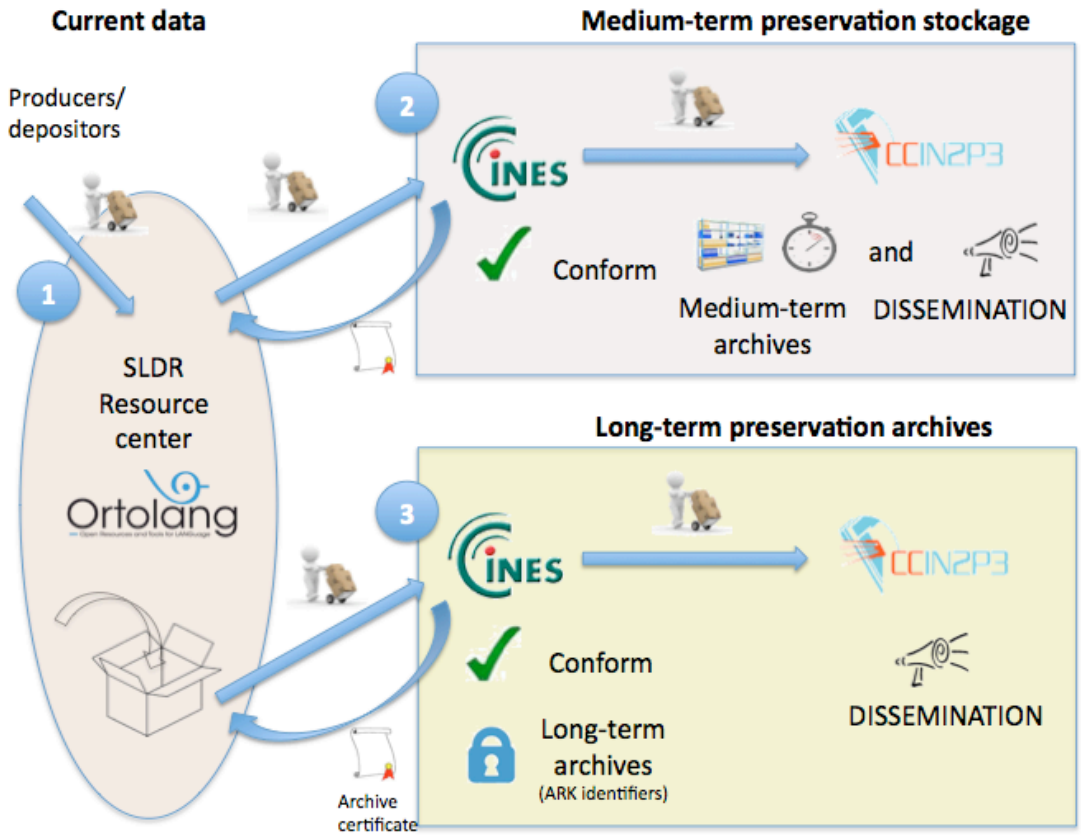

Figure 1. A descriptive outline of the data flow for medium or long-term preservation between SLDR, CINES and CC-IN2P3.

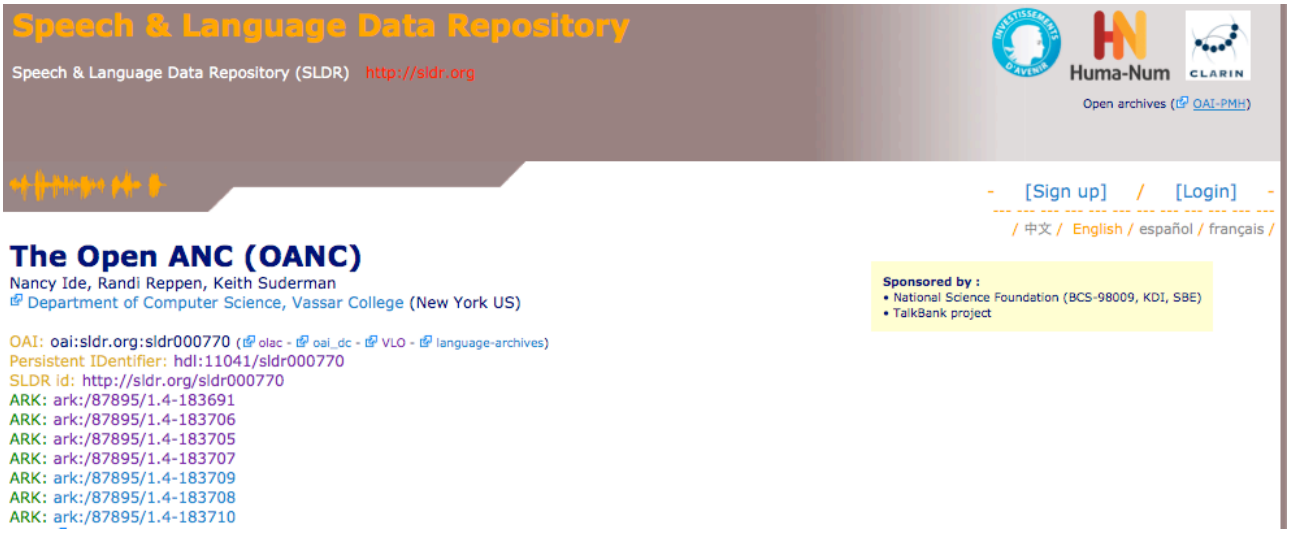

Figure 2. A descriptive page of The Open ANC (62003 files). Note the careful mention of institutional support, funding agencies, identifiers according to OAI, Handle and SLDR schemes, and Archival Resource Keys associated with the 7 segments.

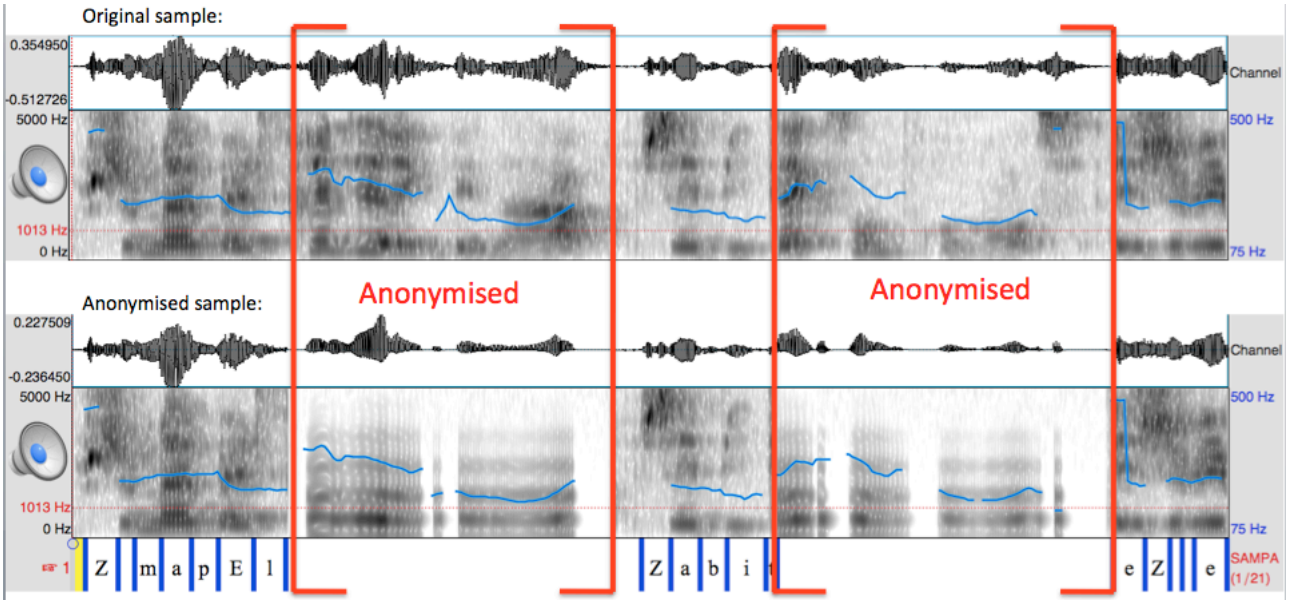

Figure 3. An illustration of the impact of the anonymisation PRAAT script by Daniel Hirst on a sentence before and after anonymisation. The pitch line is preserved.

sldr000027 – Videos of CID

Downloaded (58) primary data (corpus) Videos of CID - hdl:11041/sldr000027

First name and last name	Institution	Field of research	Date of download
M Pascal NOCERA Contact	Centre d'enseignement et de recherche en Informatique - EA 4128 (LIA, Avignon FR)	Traitement Automatique de la Parole	2008-10-17 licence #1
M Jean-Claude MARTIN Contact	Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur - UPR 3251 (Limsi, Orsay FR)	communication multimodale	2008-10-20 licence #1
M Paul ISAMBERT Contact	Langues, textes, traitements informatiques, cognition - UMR 8094 (LaTTiCe, Paris FR)	Structure du discours	2008-11-04 licence #1
M Olivier ROUCHON Contact	CINES 950 Rue de Saint Priest 34097 Montpellier Cedex 5 http://www.cines.fr/	Archivage pérenne de documents électroniques	2008-11-20 licence #1
Mme Sophie JAOUL Contact	Formes et représentations en linguistique et littérature - EA 3816 (FoReLL, Poitiers FR)	didactique	2008-12-11 licence #1

Figure 4. A excerpt from the users' community for CID videos (hdl.handle.net/11041/sldr000027)

instance sldr.org/sldr000027/metadata/olac for a description of the Corpus of Interactional Data (hdl.handle.net/11041/sldr000027). Elements such as links and table of contents are automatically computed to describe the content.

4.2.2. CMDI

In addition to Dublin Core OLAC, SLDR and its partners in the ORTOLANG project will support Component Metadata (CMDI, www.clarin.eu/cmdi). A minimum set will use the information encapsulated in DC OLAC. Later, data producers will be given the option to select one among CMDI profiles for which web forms will be available.

4.2.3. EAD, METS

Structural metadata describing the tree-structure of an item will technical details derived from file analysis will be automatically incorporated into metadata files in the EAD and/or METS format. Access to this information will be required for processing queries over large sets of data.

4.2.4. Importing metadata formats

In the long term, the repository will be able to import metadata in various formats. Import from Dublin Core, CMDI and IMDI (www.clarin.eu/imdi) is easy to figure out. Other formats (such as DDI used by social scientists) will require the mapping of elements following schemes created by the research community.

4.3. Anonymisation

Today, the strong demand for sharing linguistic data results in an increasing need for anonymising audio files for both legal and ethical reasons. This curation task requires technical support. A PRAAT script for anonymising speech recordings is distributed on SLDR (hdl.handle.net/11041/sldr000526) with particular interest for speech prosodists. It replaces segments labeled with a key word on the accompanying TextGrid with a hum sound with the same prosodic characteristics as the original sound (Figure 3).

Producing the TextGrid from a simple text file is further facilitated by the Table2TextGrid script (hdl.handle.net/11041/sldr000811).

4.4. Persistent identifiers

A necessary feature for the proper reuse of research data is the assignment of links to documents that do not depend on their location in the repository. This location is bound to vary during the life cycle of an item. First it is available as 'source data' on the submission site (SLDR), but later it is transferred to the dissemination site (CC-IN2P3) as a 'secure backup'. Once the item has become stable, producers may decide to submit it for long-term preservation, which results in yet another location.

End users and web designers expect that the URLs pointing at files or 'datastreams' remain unchanged despite these changes of location. This is accomplished by the assignment of a persistent identifier (PID) to each individual document (www.sldr.org/wiki/Handle). Referring to PIDs makes it easy for outsiders to feed their web pages or blog articles with material directly extracted from the SLDR repository.

The SLDR algorithm for assigning PIDs relies on the assumption that a document shall retain its PID across changes of its location as well as new versions of the item which it belongs to. To this effect, the identity of a file is assessed by checking its digital signature (MD5): as long as both file name and digital signature remain unchanged, the same PID is assigned. In this way, every document deposited on SLDR is accessible to automated queries regardless of its archival status.

5. Accessing research data

There is a genuine interest for the interoperability of repositories for Digital Humanities, here meaning the possibility of launching analytical processes over sets of data stored in multiple repositories. Nonetheless, most current 'showcases' only work with open-accessible material.

5.1. Controlled access

Research scholars submitting data to a digital repository strongly insist on keeping control over its dissemination and access protocols. They often find it difficult to formalize access rights in limited technical frameworks set up by engineers. Admittedly, classical solutions fail to comply with the details of regulations for the protection of private data and intellectual property. This calls for an integrated management of access rights covering the broadest diversity of cases.

France takes advantage from a significant advance on archive law, namely its *Code du patrimoine* (the Heritage Code) clarifying the notion of 'public archive' with a set of formal rules regulating access to archived documents (Act of 15 July 2008, articles L213 1-5). This framework prompted a radical change of policy as any public archive is expected to be open-accessible, with the exception of 24 derogations applicable to certain categories of documents (www.sldr.org/wiki/table_derogations_en). Each derogation has been assigned a code facilitating a systematic management. A frequent derogation case is the protection of privacy (50 years, code AR048, art. L213-2, I, 3). For a recorded audio/video corpus, this derogation may be invoked to restrict access until authorisations have been signed by all participants.

In SLDR, users are assigned categories according to profiles defined by institutional producers. If no profile is available, the default SLDR profile (sldr.org/wiki/Groupes) is applied which makes a distinction between 'academics' — teachers, students and research scholars working on subjects related with linguistics — and other users, including the ones affiliated with the speech industry. This status is carefully verified at the time of signing up. SLDR default categories and procedures are similar to the 'URCS protocol roles' used by ELAR (Nathan 2013: 6).

Users granted access to restricted material are requested to check the SLDR licence (sldr.org/wiki/Licences_en) for a licit use of the resource. Peer-to-peer exchange is only allowed for items bearing a Creative Commons licence. Data producers may optionally impose additional clauses in a specific licence.

SLDR keeps records of all controlled-access downloadings. Thus, any user of a resource may consult the list of other users, check their credentials or contact them to seek information about their planned usage of the material (see Figure 4).

5.2. Shared licences

Current development of access rights management at SLDR is aiming at a social networking approach inspired by ELAR's 'protocol' approach, here meaning 'the concepts and processes that apply to the formulation and implementation of language speakers' rights and sensitivities, and the consequent controlled access to materials.' [7, p. 4].

Transactions with groups of users are facilitated by 'shared licences' granted to sets of archived items, persons or institutions. This technique applies to individuals or groups belonging to a particular community of research participants.

An example of non-commercial licence is the Buckeye Corpus of Conversational Speech distributed by Ohio State University (hdl.handle.net/11041/sldr000776). This corpus is under a licence shared by all members of a laboratory. Thus, access is granted to persons whose affiliation with the licenced institution has been authenticated by SLDR.

Shared licences may also be used for disseminating material purchased by a group of laboratories, as is the case with the Treebank collection acquired by CNRS (hdl.handle.net/11041/ldc000828).

5.3. 'Commercial' versus 'academic'

The distinction between 'commercial' and 'non-commercial' is not a rigid one. Agencies distributing language resources (such as ELDA and the LDC) adopt a pragmatic approach with respect to financial participation: scholars and public laboratories are granted access to resources at rates significantly lower than the speech industry; the resource may even been given free on request.

SLDR has provision for links with ELRA and LDC resources of this type. See for example two instances of the EUROM collection (hdl.handle.net/11041/sldr000034, hdl.handle.net/11041/sldr000035) and the Open ANC (hdl.handle.net/11041/sldr000770).

This pragmatism is not resented as discriminative by corporations because they prefer to pay a high fee for the service which implies a contractual protection against litigation. Engineers feel reluctant to use resources labelled 'public domain' because of the trouble their company might face if this free licencing is challenged due to their use in a commercial product.

5.4. Individual permission

Resource producers sometimes need to retain full control of the sharing of their material. This may be the case with scholars using SLDR for a restricted sharing of their speech corpus until the completion of their research work. In this case, owners of the resource receive mail queries sending them to a web form for granting or denying access to applicants during a given period of time.

6. New perspectives

SLDR is a promising work environment for archive curators in charge of linguistic/oral/multimodal research material. Current focus is on the automation of curation tasks such as the production of accurate metadata and the packaging of generic items following the OAIS model. In the context of ORTOLANG, new features are being integrated such as:

- Tools for automatic phonetic annotations and alignment of transcriptions (SPPAS, www.sldr.org/sldr000800);
- A systematic pre-processing of sound files including their segmentation to intonation units and MOMEL/INTSINT labeling; these will be applicable to multitrack recordings (more than 2 microphones);
- Automatic conversion of descriptive metadata and the production of structural metadata.

This list is not exhaustive. We expect that dealing with a great diversity of linguistic material from a wide range of disciplines will encourage the development of tools and procedures coping with the requirements of high-quality research.

7. References

[1] Bel, B. and Blache, P., "Le Centre de Ressources pour la Description de l'Oral (CRDO)", Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA), 25:13-18, 2006. Online: http://hal.archives-ouvertes.fr/hal-00142931 accessed on 19 Jun 2013.

[2] Bird, S. and Simons, G., "Seven Dimensions of Portability for Language Documentation and Description", Language, 29:557–582, 2003. Online: arXiv:cs/0204020 accessed on 19 Jun 2013.

[3] CCSDS, "Reference Model for an Open Archival Information System (OAIS)", Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1, August 2009.

[4] Hedstrom, M., "Digital Data Curation – Workforce demand and educational needs for digital data curators", Proceedings of conference Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnals, Florence, 11-12 December 2012 (in press). Online: http://www.rinascimento-digitale.it/conference2012/paper_ic_2012/hedstrom_paper.pdf accessed on 19 Jun 2013.

[5] Michailovsky, B., Michaud, A. and Guillaume, S., "A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive", International Conference on Speech Database and Assessments (Oriental COCOSDA), Hsinchu: Taiwan, 2011. Online: http://halshs.archives-ouvertes.fr/halshs-00620893 accessed on 19 Jun 2013.

[6] Moulaison, H.L. and Corrado, E.M., "LAM education for digital curation: A North American perspective", Proceedings of conference Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnals, Florence, 11-12 December 2012 (in press). Online: http://www.rinascimento-digitale.it/conference2012/paper_ic_2012/moulaison_paper.pdf accessed on 19 Jun 2013.

[7] Nathan, D., "Digital archiving", in P.K. Austin and J. Sallabank, [Eds], The Cambridge Handbook of Endangered Languages, 255–273, Cambridge University Press, 2001.