Implementing the OAIS for oral/linguistic resources: the Speech and Language Data Repository venture

Bernard Bel

Implementing the OAIS for oral/linguistic resources: the Speech & Language Data Repository venture

Bernard Bel
Laboratoire Parole et Langage (LPL) - UMR 7309 CNRS/Aix-Marseille University
5 avenue Pasteur, 13100 Aix-en-Provence (France)
E-mail: bernard.bel@lpl-aix.fr

Abstract

In 2008, a pilot project initiated by TGE Adonis, a large research infrastructure, brought together designers of data repositories, archivists and system engineers to set up collaborative oral/linguistic resource centres in France. This paper discusses challenging issues addressed by this team when implementing an Open Archival Information System (OAIS) bundled with an institutional archive. After the completion of the pilot project, the Speech & Language Data Repository (SLDR) underwent development for the systematic management of access rights in compliance with the French Heritage code. Its framework claims to be applicable to other systems worldwide, which would facilitate interoperability between protected repositories equipped with transfer of authentication techniques (Single Sign-On).

1. Historical background

In 2006 the office of the social science and humanities department at the French *Centre national de la recherche scientifique* (CNRS, www.cnrs.fr) issued a call for projects aiming at the creation of digital data repositories for speech research. This initiative was driven by 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 proposal could initially be perceived as a public research facility competing with a business model unfit for academic work in small units. Later on, this feeling of competition vanished thanks to the complementarity of these models (see infra).

Two projects were selected under the label *"Centre de Ressources pour la Description de l'Oral"* (CRDO), CRDO-Aix and CRDO-Paris respectively supported by *Laboratoire parole & langage* (LPL) and *Langues et civilisations à tradition orale* (LACITO). While CRDO-Paris mostly replicated the design of the existing LACITO archive (Michailovsky et al., 2011), CRDO-Aix was built from scratch after a comparative study of existing data repositories (Bel & Blache, 2006).

In 2008, a large research infrastructure (*Très Grand Équipement Adonis*, www.tge-adonis.fr) decided to promote the long-term preservation and sharing of oral resources in social sciences and the humanities. Benoît Habert, a professor of linguistics and the joint director of TGE Adonis, suggested that a pilot project be dedicated to oral/linguistic resources. Claude Huc, a founder of the PIN group (*Pérennisation de l'Information Numérique*, www.pin.association-aristote.fr/doku.php) in charge of digital archiving at the *Centre national d'études spatiales* (CNES, www.cnes.fr), was appointed to coordinate this pilot project.[1]

The project infrastructure involved the two branches of CRDO as submission sites connected with two major computing centres, *Centre informatique national de l'enseignement supérieur* (CINES, www.cines.fr) and *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3, cc.in2p3.fr), respectively the archiving and dissemination sites (Bärring, 2008). Following the example of spatial agencies, managers opted for the Open Archival Information System (OAIS) promoted by the *Consultative Committee for Space Data Systems* (CCSDS, 2011) and normalised in 2003.

The objective of the pilot project was threefold:

1) To implement the OAIS information and functional models taking into account specific requirements of speech data in a framework that should be generic enough to encompass the diversity of data structures and work environments of social sciences and humanities;

2) To produce persistent archives owing to the project's association with CINES, a national institution commissioned by the Ministry of Higher Education and Research for the long-term preservation of digital information.[2]

3) To facilitate the dissemination of oral/linguistic resources via the TGE Adonis grid at CC-IN2P3 giving access to 'datastreams' on the Fedora Commons

[1] Project documents, reports and evaluation are stored in the archive: www.sldr.org/ark:/87895/1.4-187408

[2] Legal statements were signed in June 2010 between the CNRS, CINES and DAF (the French National Archive) after the evaluation (Marcoux, 2009) and presentation of the pilot project (Pouyllau, 2009).

platform.

Issues specific to speech data and their descriptive metadata were addressed, with preference for solutions extendable to data types outside the field of linguistics. Access to data could be restricted in compliance with *Code du patrimoine* (the French Heritage code) in its July 2008 version.

CRDO-Paris and CRDO-Aix became fully operational in Summer 2010 as their long-term preservation modules switched to the 'production' mode. However, CRDO-Aix continued using the 'test' mode as a platform for medium-term preservation, an important feature for the incremental construction of items before the completion of research/documentation projects.

At the term of this experimental phase, partners were instructed by INSHS (the social science and humanities institute of CNRS) to give up the 'CRDO' acronym. Thereafter, CRDO-Aix was renamed *Speech & Language Data Repository* (SLDR, www.sldr.org) without modifying its operational process. CRDO-Paris was renamed CoCoON (cocoon.huma-num.fr).

In September 2011, the SLDR dealing with oral data, and the CNRTL (*Centre National de Ressources Textuelles et Lexicales*, managed by ATILF in Nancy) dealing with written data, became the points of anchoring of project ORTOLANG which aims at implementing in France a network of CLARIN centres dedicated to linguistics.[3]

2. Our OAIS implementation

The functional diagram describing data flow in our OAIS implementation (see figure 1) is similar to the one borrowed from the CNES spatial agency.

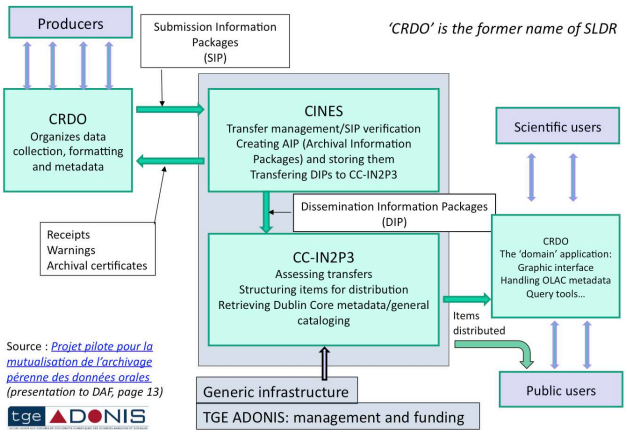

Figure 1: Data flow in the pilot project, translated from: www.sldr.org/docs/admin/AP-ADONIS_vers_DAF.pdf

The archive site (CINES) receives *Submission Information Packages* (SIP) from its submission sites (CRDO-Paris and CRDO-Aix, currently SLDR). Each package contains an *item* of generic type bundling together digital documents, in other words a tree-hierarchy of files and folders with no restrictions on depth and size.[4] The distinction between item types (primary data, secondary data and tools in SLDR) is therefore the matter of (domain-specific) descriptive metadata.

SIPs contain 'archive-compliant' data eligible for long-term preservation, thereby meaning formats chosen in a list of acceptable standards (see infra). Every valid SIP is stored as an *Archival Information Package* (AIP) in the archive (managed by the Arcsys software). In the same time, the archive site creates a *Dissemination Information Package* (DIP) transferred (via iRods) to the dissemination site (CC-IN2P3), as shown figure 2.

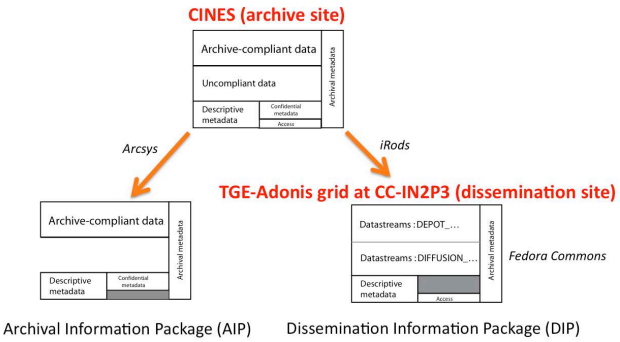

Figure 2: Management of a SIP at the CINES site.

The contents of SIP, AIP and DIP are similar expect for the fact that SIPs and DIPs may carry data that is not eligible for long-term preservation. This requires an explanation as this feature had not been implemented in the initial model. In an institutional archive (such as CINES) it is the task of archive managers to make sure that digital documents are preserved both in their

[3] ORTOLANG (www.ortolang.fr) is a beneficiary of French State support via its "Investissements d'avenir" programme (ANR-11-EQPX-0032).

[4] More details on www.sldr.org/wiki/Packaging-en

physical integrity and readability. To this effect, CINES has the commitment to migrate data to new formats whenever necessary. This results in a limitation of acceptable formats on the CINES platform. Archive-compliant formats must be non-proprietary, widely used and fully documented.[5] Since the current list of acceptable formats[6] does not entirely fit with practice of oral/linguistic resources, each SIP is uploaded along with a specific folder (named 'DIFFUSION') whose content is ignored by the archiving system and forwarded to CC-IN2P3 once the SIP has been validated. The 'DIFFUSION' folder may therefore contain files in formats that are not accepted by the archive but accessible to users.[7]

DIPs received by CC-IN2P3 are ingested as 'datastreams' on its Fedora Commons platform designed for efficient storage and dissemination.

A flow chart summarizing the entire submission process is shown figure 3.[8] This process is identical for long-term and medium-term preservation, the only difference being that in medium-term preservation the archive site does not store AIPs.

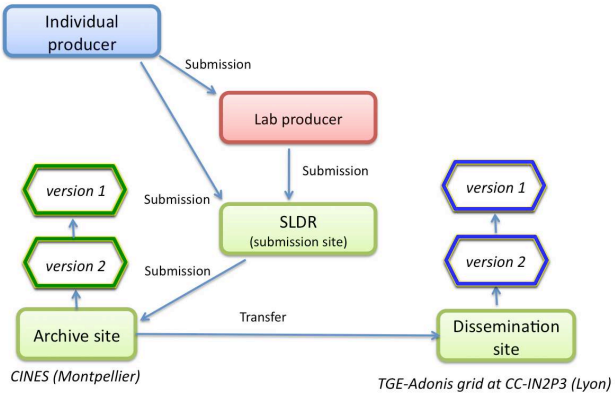

Figure 3: SLDR submission process

3. Technical limitations of the model

The first limitation of SIP formats is on file names and hierarchy. Both the archive software and Fedora Commons have limitations on file name encodings and lengths. In addition, datastreams are stored 'flat' at the tree root.

For these reasons, file names are serially tokenized by the submission site, e.g. '341.wav', '342.pdf' etc.[9]

A mapping of original file names with their datastreams is stored among documentary files (as a plain text UTF8 file). This mapping also contains parameters (file size, path in the hierarchy, modification time etc.) facilitating the safe reconstruction of source items from their dissemination datastreams, as shown figure 4.

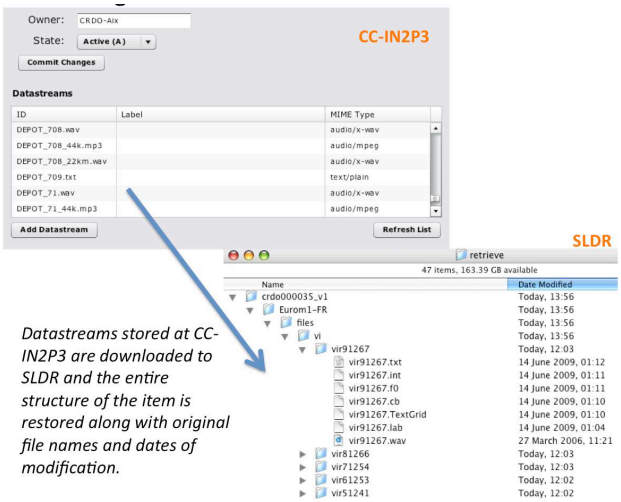

Figure 4: Retrieving an item from its datastreams stored at the dissemination site

Another limitation is the size of a DIP for its ingestion by the Fedora Commons platform. There is a physical limit (approx. 32,000) in the number of datastreams which has an even more dramatic incidence on the speed of ingestion. For this reason we set up limits of 30 Gbytes and 10,000 files for the submission package. Subsequently, bigger packages are automatically segmented.[10]

4. Access to items

Controlled access to items is described on figure 5. The user clicks a link sending a query to the SLDR following links displayed in descriptive metadata. If the user signs on as an authorised person for this particular item and agrees with the SLDR licence, the system forwards the query to the dissemination site. The site replies opening a 'channel' allowing the user to download files for 24 hours. The user receives an email making it easy to resume downloading, if interrupted, without going again through the authentication process.

[5] This compliance may be checked using CINES page: facile.cines.fr. Addition of new file formats is negotiated between CINES and data producers.

[6] www.sldr.org/wiki/Formats

[7] A typical example is a Skype dialogue recorded as a QuickTime™ movie with four sound tracks (two for each speaker), which is suitable for speech analysis. The archive-compliant version of this movie, according to CINES specifications, would be a MP4 video with a stereo mix of the four tracks in AAC format, along with two stereo WAV or AIFF files containing the sound tracks in their original encoding.

[8] In our implementation, the archive and dissemination sites are located in remote cities and managed by different institutions. This configuration, which requires an efficient team coordination for its maintenance, is not compulsory. Site locations could be physically identical provided that remote-distance backups are performed.

[9] This tokenization works in a different manner for files contained in a 'SECRET' folder. Read: www.sldr.org/wiki/Packaging-en

[10] For example, the *Open ANC* (www.sldr.org/sldr000770) which contains more than 60,000 XML files has been chunked to 7 segments, an operation that remains invisible to end users. Nonetheless, the ingestion of this corpus by the Fedora Commons platform took about three days!

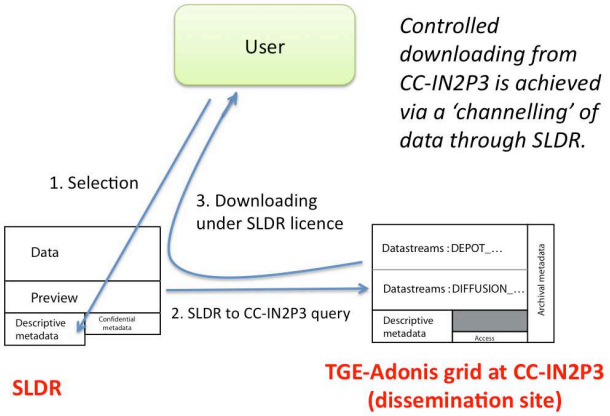

Figure 5: Controlled downloading from the SLDR

Another process is illustrated figure 6 for datastreams declared in open-access. These may be downloaded directly from the dissemination site if their URLs have been recorded.

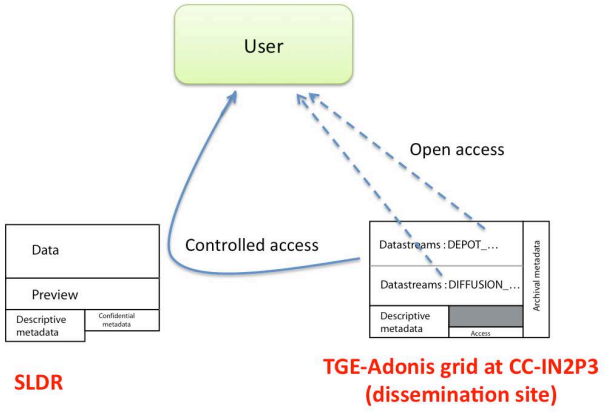

Figure 6: Two access modes on the Speech & Language Data Repository

It should be noted that the archive site (CINES) is never involved in disseminating archived material. Its only commitment is to resend DIPs to the dissemination site in case of a failure of the site.

Data is available from CC-IN2P3 rather than being downloaded from the submission site (SLDR). Since the original item can be safely rebuilt from its datastreams (see supra) it is possible to delete source material from the submission site. Space allocation, therefore, is only the care of CINES and CC-IN2P3.[11]

5. Metadata and descriptive files

Metadata associated with items contain three types of information:

1) Descriptive metadata (domain-specific). Though there was no obligation on the part of CINES we opted for Dublin Core with OLAC qualification for the minimum description of items.[12]

2) Archival metadata for the CINES platform (generic). These contain a minimal description of the item using a few (unqualified) Dublin Core elements to facilitate item identification in the long term.[13]

3) Dissemination metadata. This feature is under study at CINES, taking into account the recent formalisation of access rights procedures according to the French *Code du patrimoine* (see infra). The Speech & Language Data Repository anticipated this process by creating 'accessRights.xml' files which producers can edit for assigning access parameters to any directory.[14] These files are preserved as documentary files and forwarded to the dissemination site.

6. Relations

At an early stage of the pilot project, the archive site was set up for the storage of individual information packages, e.g. sets of publications or PhD memoirs. Oral/linguistic resources introduced the need for linking items through relations spelled out in Dublin Core. We had long discussions to decide on a minimum description that would take some of these relations into account without introducing domain-specific categories. Finally two types of relations were standardised: *filiation* and *maj* (*'mise-à-jour'* = 'update'). The latter provides a link to the first version of an item. The former is a link covering any of the following Dublin Core relations: *isPartOf, isRequiredBy, isVersionOf, replaces* and *references*.[15]

7. Implications of long-term preservation

In their conclusive paper on the pilot project experiment (2011), Habert and Huc raised the issue of 'archiving' in the context of scientific research as examplified by linguistics:

Somehow, in French, the very word 'archives' has a "dusty" connotation. It reminds of long forgotten books, papers or even manuscripts, which are accessed now and then only for history researches. It is not the case in English for 'archive' (see the role of arXiv) or 'repository': as a matter of fact, these words, at least for digital contents, do not imply long-term preservation, but

[11] In practice, only items stored as stable versions in long-term preservation are deleted from the submission site. Items subject to changes and versioning are maintained on the site. This source material can be downloaded by the item owner and designated authorised users. This makes it easy for a team to share up-to-date data instead of uploading a new version each time a change has been performed.

[12] In the context of our association with CLARIN (ORTOLANG project) we are in the process of adding Component Metadata (CMDI, www.clarin.eu/cmdi) using and categories handeled by ISOcat (www.isocat.org/files/12620.html). Next steps will be the implementation of TEI for annotations and RDF for structured metadata.

[13] For instance, wheras the name of an item of SLDR may be harvested in up to 5 language versions from its OLAC metadata, all versions are copied to unqualified elements such as:
<title>en: Renivier's language, Vanuatu, Malekula</title>
<title>es: La lenguaje de Renivier, Vanuatu, Malekula</title>
<title>fr: La langue de Renivier, Vanuatu, Malekula</title>

[14] For more details, read: *Access rights settings*. www.sldr.org/wiki/accessRightsSettings_en

[15] These simple relations are stored in the 'sip.xml' file supplied with each version of the item.

rather focus on easy access. Anyway, in SSH [Social Sciences and Humanities] *a long-term preservation infrastructure needs to ensure that it does not deliver a "still life" of the research. On the contrary, current research must be able to use past research, as secondary data for instance. New analyses of primary data must be linked to it, in order for them to be falsifiable. It is often necessary to mend an annotation (when transcription conventions are updated, for instance) or to change metadata (for instance when access rights change). New corpora must be made available as soon as possible. Therefore archived data are living, ever evolving material.*

These statements summarise critical issues in applying the OAIS model to oral/linguistic resources. Most problematic, long-term preservation in an institutional archive implies that every single version of an item shall be preserved forever. In the original model borrowed from spatial agencies, any change in the content or description of a document would result in the storage of a new version of the entire item. Consider for instance the case of a video corpus sized to 200 Gbytes! To cope with this, the model was refined to handle two parallel versioning processes: each version of the item generates a new AIP with its own persistent identifier (ARK[16]) containing data, descriptive metadata and documentary files. However it remains possible to upgrade the contents of descriptive metadata and documentary files in the same version; these upgrades are assigned their own ARK identifiers and they are linked to the original package via a specific relation.

This is fair practice since primary data (sound/video files) are generally unchanged whereas documentary files and metadata are likely to be updated.[17]

8. Legal constraints

A legal constraint associated with the storage of resources in an institutional archive is that they qualify as 'public archives'. The French *Code du patrimoine* (Heritage code, L211-4) states that public archives are *"documents produced by the activity of State, local governments, public institutions and other legal persons under public or private law who are in charge of a public service, as part of their public service remit."*[18] Most language resources collected for their patrimonial value and/or usability by research scholars connected with public institutions are eligible for this qualification.

Another statement (L213-1) stipulates that *"public archives shall be immediately in open access, unless subject to restrictions as per article L213-2."* This is a innovation because in earlier versions (before 15 July 2008) a delay of 30 years was granted for the dissemination of public archives. Nonetheless, article L213-2 gives provision for 25 cases of derogations to this open-access obligation. These derogations have been encoded and tokenized to set the ground for a systematic management of access rights.[19]

In addition, the open-access obligation only refers to the informational content of a resource from a patrimonial viewpoint. For instance, displaying the entire content of a sound/video recording via a streaming device does satisfy this obligation. Therefore it is not illicit to restrict the downloading of 'high-resolution' files to users belonging to the academic community or a laboratory authorised via shared licences. This regulated dissemination implies the acknowledgement of SLDR non-commercial licence[20] and a trace of downloadings that is visible to all users of the same item.[21]

A frequent derogation case is the protection of privacy (50 years, code AR048, L213-2, I, 3). For a recorded audio/video corpus we consider that this derogation is granted by default until authorisations have been signed by all participants. These must comply with the ethics of informed consent, making it clear that the participant is aware of the type and range of dissemination planned for the document. Participants may also decide that a particular audio/video recording is worth sharing with scholars though it shall not be displayed in a public presentation. In this case, the standard SLDR licence signed by users will be completed with an additional licence.[22] The text of this licence is submitted to participants as an annex of their consent form.[23]

9. A systematic management of access rights

The new French legal framework for public archives is altogether a significant progress for scholars and a headache for archive curators because of the sudden change of status of documents scattered in their collections. Having anticipated this difficulty during our initial discussions with archivists and lawyers, we figured out a machine-friendly model for the storage of dissemination metadata associated to items, folders and individual files.[24] This model has been favorably appreciated by archive curators and computer scientists in charge of scientific data repositories (Bel, 2011; 2012a-b-c).

---

[16] *Archival Resource Key*
www.cdlib.org/inside/diglib/ark/arkspec.html

[17] As long as primary data is not stable (when new recordings are being produced and added to the package) items are stored in medium-term preservation. In this mode, new versions can be piled up since all of them will be deleted at the time the item is sent to long-term preservation.

[18] www.sldr.org/wiki/CodePatrimoine

[19] This table has been translated to all navigation languages of SLDR. The English version is available from www.sldr.org/wiki/table_derogations_en

[20] www.sldr.org/licence/en

[21] Users' communities, see www.sldr.org/com

[22] For example www.sldr.org/sldr000761/licence/LicenceStRemy.pdf

[23] A complete consent form taking this additional licence into account is proposed here: www.sldr.org/doc/forms/ConsentementModele_en.pdf

[24] *Access rights management in compliance with the French Code du patrimoine: a generic approach for the OAIS model run by SLDR*.
www.sldr.org/wiki/accessRightsManagement_en

Article L213-5 of *Code du patrimoine* states: *"Any Administration holding public or private archives is required to give reasons for denying access to archival documents."* In practice, this means that if a user clicks the *"List of downloadable files"* link on an item,[25] moving the mouse over a link displays the legal status of its content (in the navigation language): the reason for denying access along with the date at which it will fall into the public domain. as shown figure 7.

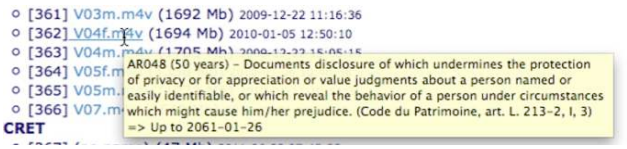

Figure 7: Denying access to file 'V04f.m4v'

Links in open access are tagged by a special icon and the mouse-over notice justifies file accessibility, as shown figure 8.

Figure 8: 'Master2AudreyThomas.pdf' is in open access.

It is obvious that archive curators cannot maintain an agenda of documents whose status is bound to change after decades… Therefore, this status is automatically taken care of by the SLDR. For each item the date of the ealiest deadline for modification is stored in the database. Everyday the system looks for items containing at least one document whose status needs to be modified. The administrator receives a list of these items.

Displaying the map of documents in an item reveals their public/private status. If the status is compliant with the Law it is displayed in green; otherwise it appears in red. An administrator can quickly edit the status of the concerned document. The system then prepares a 'metadata update' SIP that will be sent to CINES in the next connection. Once this update has been forwarded to CC-IN2P3 it results in modifying access to the concerned document(s).

The same process applies to documents whose access status needs to be turned to 'private' as their authorisation reached a date of expiry, or participants changed their minds regarding open access. Access rights may also need to be revised in earlier versions of an item since datastreams earlier declared open remain accessible.

The model of access rights management currently implemented in the Speech & Language Data Repository applies to the French legal system governing its associated institutional archive. However its design is flexible enough to anticipate evolutions of the system or create variants applicable to other systems worldwide. Interoperability should thus be facilitated for protected repositories equipped with transfer of authentication techniques (Single Sign-On).[26]

10. Shared licences

The Speech & Language Data Repository is bound to play a role in the preservation and sharing of linguistic resources produced by academic institutions because of its secure reliance on an institutional archive and the flexibility of its management of access rights. To anticipate this process it was necessary to figure out a method for a controlled dissemination of resources adapted to the specific needs of their producers.

Consider for instance the *Buckeye Corpus of Conversational Speech* distributed by Ohio State University.[27] This corpus is under a non-commercial licence that may be shared among the members of a laboratory.[28] If we receive evidence that an agreement has been signed between a laboratory and the OSU for the sharing of this corpus, we will store the receipt and assign an authorisation tag to the laboratory. Consequently, any person identified as a member of this lab will be granted access to the SLDR copy. This technique of shared licences may be extended to individuals selected by the rights holder of a resource.[29]

11. 'Commercial' versus 'academic'?

In theory, a resource aimed at commercial distribution is not eligible for long-term preservation (as a 'public archive') on the CINES archive site. However the distinction between 'commercial' and 'non-commercial' is not a rigid one. We already argued that a document may be declared 'open-access' even though access to its 'high-resolution' version is reserved to certain categories of users. In a similar way, agencies distributing language resources (such as ELDA and the LDC) adopt a pragmatic approach with respect to financial participation: scholars and public laboratories may be offered resources at rates much lower than the speech industry; in some cases the resource is even given free on request.[30] This distinction may also be spelled out with a complete version of the resource available on a paid basis and a smaller version freely available.[31]

[25] Or follow a link such as: www.sldr.org/sldr000764/toc

[26] To take steps towards a standardisation, we are keen to receive feedback on this initial work and collaborate with the design teams of data repositories.

[27] www.sldr.org/sldr000776

[28] www.buckeyecorpus.osu.edu/php/faq.php

[29] Details on www.sldr.org/wiki/SharedLicence

[30] This pragmatism is not resented as discriminative by professionals of the private sector because companies prefer to pay a full fee for the service provided as this payment implies a contractual protection against litigation. Private sector engineers feel reluctant to use resources declared 'public domain' because of the trouble they might face in case this free availability is challenged due to their use in commercial products.

[31] SLDR has provision for links with ELRA and LDC resources of this type. See for instance: www.sldr.org/sldr000034, www.sldr.org/sldr000035 (ELRA)

12. Conclusive remarks

As put by Habert & Huc (2010:426),

"In attempting to build a lasting archiving infrastructure, the main difficulty is building shared representation between all the actors who are involved. They need to agree on the way the data and the metadata are organized, on how it is going to be accessed and used. Even more crucially, the overall process and the precise division of responsibilities must be agreed upon. […] *What is at stake is not "implementing the OAIS model", but finding together a possible meaning for it in a specific context. In the end, the solution will be reliable if and only if a deep agreement is obtained on the overall scheme as well on the detailed procedures."*

The coordinators of TGE Adonis/CRDO/CINES/CC-IN2P3 pilot project were successful in promoting the spirit of team work and taking the project to its completion.[32] The author feels indebted to colleagues who dedicated their time and expertise to this project: Pascal Calvat, Pascal Dugénie, Stéphanie Girault, Benoît Habert, Claude Huc, Michel Jacobson, Pierre-Yves Jallud, Thomas Kachelhoffer, Nicolas Larrousse, Olivier Rouchon and Huân Thebault.

We further hope that national policy makers will maintain the construction of large research infrastructures for social sciences and humanities in synergy with networks such as CLARIN [33] and DARIAH[34] in Europe.[35] This commitment is crucial at a time the international scientific community is in strong demand for cooperative resource development and sharing.

13. References

Bärring, O. (2008). Hosting of IT services and data for Human and Social Sciences in France. A preliminary study for TGE Adonis (Contract Nr K1432). sldr.org/docs/admin/RapportBarring.pdf

Bel, B. (2011). A presentation of the OAIS model and the systematic management of access rights. CLARIN-D tutorial. MPI for Psycholinguistics (September 7-9: Nijmegen, The Netherlands). sldr.org/doc/show/NijmegenPresentation.pdf

Bel, B. (2012a). Mutualisation et archivage pérenne des données orales : un nouveau cadre technique et juridique au service de la recherche en linguistique. Colloque *Les Archives de la Recherche : problèmes et enjeux de la construction du savoir scientifique* (January 18-20: Paris, France). www.lpl-aix.fr/article/4871

Bel, B. (2012b). A Trusted Digital Repository based on the OAIS model with integrated management of access rights. Proceedings of *Cultural Heritage On Line - Trusted Digital Repositories & Trusted Professionals* (December 11-12: Florence, Italy), pp. 1--6. hal.archives-ouvertes.fr/hal-00983703

Bel, B. (2012c). Une implémentation du modèle OAIS pendant (et après) le projet pilote TGE Adonis/CINES/CC-IN2P3/CRDO. Journées OAIS (2012 octobre 23-24 : Lyon, France). sldr.org/doc/show/lyon-oct2012/UneImplementationOAIS.pdf

Bel, B.; Blache, P. (2006). Le Centre de Ressources pour la Description de l'Oral (CRDO). *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence* (TIPA), 25, pp. 13--18. hal.archives-ouvertes.fr/hal-00142931

CCSDS (2009). Reference Model for an Open Archival Information System (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August. public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx

Habert, B.; Huc, C. (2010). Building together digital archives for research in social sciences and humanities. Social Science Information 49, 3, pp. 415--443. hal.archives-ouvertes.fr/hal-00466352_v1

Marcoux, Y. (2009). *TGE Adonis – Projet d'archivage des données produites en France par les SHS/Projet pilote sur les donneées orales, novembre 2008 – avril 2009. Rapport d'expertise sur la version préliminaire du résumé opérationnel.* sldr.org/docs/admin/Marcoux-resume-operation.pdf

Michailovsky, B.; Michaud, A.; Guillaume, S. (2011). A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive. International Conference on Speech Database and Assessments (October 26-28: Hsinchu, Taiwan, Province of China) hal.archives-ouvertes.fr/halshs-00620893_v1

Pouyllau, S. (2009). *Présentation à la Direction des Archives de France du projet pilote d'archivage pérenne des données orales.* 23 October. sldr.org/docs/admin/AP-ADONIS_vers_DAF.pdf

---

www.sldr.org/sldr000770 (LDC)

[32] Read *CRDO-Aix report on the pilot project coordinated by TGE-Adonis, January 2011.* www.sldr.org/wiki/BilanProjetJanvier2011_en
CRDO-Paris report may be found on www.sldr.org/doc/admin/bilanProjetPilote.pdf

[33] www.clarin.eu

[34] www.dariah.eu

[35] Habert & Huc (2010:430) expressed a shared concern: *"Since the 17th century at least, France has been a very centralized state. The past thirty years dramatically changed this tendency, with a law on decentralization in 1982 and another law in 2007 granting more autonomy to universities. To make a long story short, there is now a contradiction between a centralized state and centrifugal forces. The first fear is that this contradiction could be detrimental to the creation of infrastructures. The second fear concerns the real intention and/or capacity for the French State to manage in a continuous and consistent way the construction of large infrastructures for the SSH. International partnerships and huge investments protect "historical" Large research infrastructures in physics, for instance. This is not the case in SSH. In the French roadmap for Large Research Infrastructure (December 2008), the provisional budget for SSH represented 1.5% of the total budget…"*