The OAIS model in Digital Humanities: feedback from archival practice in linguistics


Bernard Bel <contact@sldr.org>
Laboratoire Parole et Langage (LPL)
CNRS – Aix-Marseille University

Tavola rotonda: Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis. Centro Linceo Interdisciplinare “Beniamino Segre”, Roma (Italy), 14 sept. 2015

ORTOLANG is a beneficiary of French State support via its “Investissements d'avenir” program (ANR–11–EQPX–0032).
Abstract

• OAIS is a conceptual framework for an Open Archival Information System (ISO 14721:2003) initially designed for constructing archives of quantitative data. In 2006, the National Center for Scientific Research in France (CNRS) launched a pilot project setting-up services for the sharing and long-term preservation of speech data and related material. Several institutional partners were involved, among which the CINES (National Computing Center for Higher Education).

• The Speech & Language Data Repository (SLDR, www.sldr.org) is an outcome of this project. Success of the venture depends on fostering an effective collaboration between the major actors: research scholars, curators, archivists and computer scientists. We will expose (positive and negative) examples of digital curation to identify good practice in terms of deposit, descriptive metadata, integrated access-rights management and the packaging of archival items for their medium and long-term storage.
Who are we?

The Speech & Language Data Repository (SLDR – www.sldr.org) is a digital repository for the sharing and long-term preservation of linguistic data created by Laboratoire Parole et Langage (LPL, UMR 7309 CNRS-AMU).

The aim of SLDR is to bring together resources associated with oral communication. This is necessary for avoiding the waste of time and money creating again resources that are already available. Further, resource sharing in speech research is a strategic response to the challenge of capitalizing knowledge from low-interconnected scientific domains such as descriptive, formal and computational linguistics, sociolinguistics, literature, language translation, neuroscience, psycholinguistics etc.

SLDR is open to research projects regardless of geographical/national locations. The site is accessed in four languages: English, Chinese, Spanish and French.

SLDR is in the process of migrating its data and services to the new ORTOLANG platform (Open Resources and TOols for LANGUAGE) in which new tools for the processing of oral/written linguistic data will be made available. ORTOLANG is a service centered on language studies completing the general offer of Huma-Num (a Very Large Research Infrastructure), aiming at interoperability with European networks CLARIN (linguistics) and DARIAH (Digital Humanities).

SLDR architecture is based on the Open Archival Information System (OAIS) and it guarantees access to documents via persistent identifiers. The dissemination site is hosted by INIST. Submissions for long-term preservation are sent to CINES following a procedure engineered by Huma-Num.
A broad range of disciplines...

from experimental linguistics...
... to field linguistics
... including didactics, processing written forms, understanding of speech information, the assessment and retraining of voice, speech and language dysfunctioning etc.
Items of SLDR include speech/singing corpora, their annotations, lexica and other knowledge bases as well as tools associated with data processing. Corpora may comprise audio/video recordings and measurements of physiological activity.
Why and how we use the OAIS model
A classical set-up: safe storage and dissemination
A classical set-up: safe storage and dissemination (ctd.)

Problems commonly associated with safe storage

1. Access to data is centralised: the dissemination site is also the submission site. => Week argument
2. Financial support to backups is granted for ... years. What will happen beyond this period? Who will negotiate new arrangements? On what legal/ethical grounds? => Strong argument
3. Format migration is under the care of data producers. What will happen in the long term? => Very strong argument

The OAIS solution

- Preservation + migration of data is handed over to an institutional archive (or, at least, a Trusted Data Repository).
- This solution generates technical and legal constraints.
Why and how we use the OAIS model...

- The Open Archival Information System (OAIS) is ISO 14721 standard. It was imposed by institutions heading our pilot project for the sake of formal agreements between the French National Archive, CINES and CNRS.

- We drew experience from the practice of long-term data preservation in astrophysics.

- The actual implementation takes into account specific features of oral/linguistic resources, notably:
  - The diversity of file formats: sound/video, all signals associated with speech/singing, pictures, texts, tables etc.
  - Secondary data (annotations etc.) and metadata are mutable and extensible.
  - A multilingual approach of descriptive metadata, extra-European scripts, transliteration/annotation standards (IPA etc.)
Why and how we use the OAIS model...

• Aim 1: preserving data

• Aim 2: making it accessible and eligible for reuse after an unspecified period of time. This is the challenge of long-term preservation.

• These issues deal with the very long term, practically meaning more than 30 years.

• Nonetheless, medium-term preservation is also based on the same OAIS framework. => This is cost-effective!
Implications of long-term preservation

• Preserve the **intelligibility of a document**
  
  o The archive site accepts a restricted number of **preservation formats** whose specifications are freely accessible. => [http://sldr.org/wiki/Formats](http://sldr.org/wiki/Formats)
  
  o WAV, AIFF, AAC, FLAC, MP4, OGG, MKV, PDF, PDF/A, PNG, TIFF, SVG, GIF, XML, HTML, TXT, ODT
  
  o The archive site (**CINES**) is committed to **migrating** formats once they have become obsolete. *This is the job of archive curators, not data producers!*

• Preserve the **signification of its content**
  
  1. **Descriptive** metadata (public + confidential);
  2. **Archival** metadata;
  3. A formal description of **conventions** used in the archive. PPDI = *Project Preservation Description Information.*
    
    [http://sldr.org/ppdi](http://sldr.org/ppdi)
The “pilot project”
OAIS framework
(2008-2010)
The “pilot project” OAIS framework

The pilot project involved 3 actors:

1) **Submission sites:** CRDO-Aix (SLDR) and CRDO-Paris

2) **An archive site:** Centre informatique national de l’Enseignement supérieur (**CINES**, Montpellier)

3) **A dissemination site:** the TGE-Adonis grid at Centre de calcul de l’Institut national de physique nucléaire et de physique des particules (**CC-IN2P3**, Lyon)
A functional diagram of the OAIS setup

Producers

CRDO/SLDR
Organizes data collection, formatting and metadata

Submission Information Packages (SIP)

CRDO/SLDR
The ‘domain’ application:
Graphic interface
Handling OLAC metadata
Query tools...

CINES
Transfer management/SIP verification
Creating AIP (Archival Information Packages) and storing them
Transferring DIPs to CC-IN2P3

Dissemination Information Packages (DIP)

CC-IN2P3
Assessing transfers
Structuring items for distribution
Retrieving Dublin Core metadata/general cataloging

Generic infrastructure

TGE ADONIS: management and funding

Scientific users

Public users

Receipts
Warnings
Archival certificates

Source: projet pilote pour la mutualisation de l’archivage pérenne des données orales
Medium-term preservation: Pilot project (2008-2010)

SIP = Submission Information Package
DIP = Dissemination Information Package
Long-term preservation:
Pilot project (2008-2010)

SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

(1) SLDR
(2) Submission

18 page
Two problems solved by the Pilot project...
How to deal with mutable/extendible data?

**Problem**

- In a given item there are bits of information that remain persistent (e.g. primary data of sound/video resources) whereas others (such as descriptive metadata) may need to be **updated more frequently**.
- An institutional archive preserves **all versions** of information packages that have been submitted.
- It would be detrimental to submit again the entire package (including large amounts of primary data) every time a **small amount** of descriptive data has been modified.

**Solution worked out with CINES**

- Each *Submission Information Package* (SIP) contains two folders: one which is dedicated to **persistent data** and the other one to **metadata and documentary files**.
- It is possible to submit a new version either for the entire package or for its documentary files. The latter is named ‘**metadata updating**’.
What to do with ‘uncompliant’ files?

Problem
• A few file formats (e.g. MP3, FLV) are uncompliant with long-term preservation though they are needed for data dissemination (e.g. streaming).
• In the OAIS model all data is first sent to the archive site, checked, stored in the archive and thereafter forwarded to the dissemination site.

Solution worked out with CINES
• Submission information packages contain a dissemination folder whose content is ignored by the archive site. It is forwarded to the dissemination site as part of the Dissemination Information Package (DIP).
• This dissemination folder may also contain files requiring medium-term preservation because they are subject to updates/deletion and/or their formats are not (yet) recognised by the archive site.
The “SLDR/ORTOLANG” OAIS framework
OAIS on SLDR/ORTOLANG: medium-term preservation

Producer

submit

Workspace

« Source item »

Team members

SLDR/ORTOLANG

Dissemination site

‘Snapshots’ of submitted versions

Dissemination Information Packages

Version x

Version x-i

Version 1

Users

DIP

DIP

DIP

DIP
OAIS on SLDR/ORTOLANG: long-term preservation

Producer

CINES

AIP

AIP

Vx

V1

Workspace

« Source item »

Team members

Users

Dissemination site

DIP

Version x

Version x-i

Version 1

Dissemination Information Packages

Archival Information Packages

SLDR/ORTOLANG

Submission

SIP

Vx

CINES

DIP

Version x

DIP

Version 1

(a) archival certificate

(2)
Creating submission information packages (SIP)
Structure of the *Submission Information Package*

- **SIP**
  - sip.xml
- **DEPOT**
  - Archive-compliant data
  - Metadata and documentary files (archive-compliant)
- **DIFFUSION**
  - Archive-uncompliant data

Metadata and documentary files (archive-compliant)
Creating the submission information package

(See: http://sldr.org/wiki/Packaging-en)
Archival metadata
Archival metadata: the ‘sip.xml’ file

This file contains archival metadata: a brief description of the item, relations to other items, versioning, storage instructions and specifications of all files submitted for long-term preservation.
Archival metadata: the ‘sip.xml’ file

--- SIP de crdo000759_v2 ---
Nouvelle version = 2
"<?xml version="1.0" encoding="UTF-8" ?>
http://www.cines.fr/pac/test/sip.xsd">
<DocDC>
<title>en: Gangubai</title>
<title>es: Gangubai</title>
<title>fr: Gangubai</title>
<title>mr: गंगूबाई</title>
<creator>Bernard BEL</creator>
<creator>Hema RAIRKAR</creator>
<subject>en: subaltern studies, cooperative action
fr: Etudes sur la subordination, recherche
mr: धंतिपरंपरीय आचरण, सहकारी गतिविधिक</subject>
<subject>en: An interview with Gangubai Ambore
fr: Entretiens avec Gangubai Ambore
mr: गंगूबाई अंबोरे म्हणजेच मिळालेल्या आम्ही</subject>
<publisher>Centre for Cooperative Research in Sangeet Natak
in India</publisher>
<creator>Bernard BEL (creator)</creator>
<creator>Hema RAIRKAR (creator)</creator>
<creator>Gangubai AMBORE (singer)</creator>
<creator>Jitendra MAID (interviewer)</creator>
<creator>Bhimsen NANEKAR (interviewer)</creator>
<creator>Andréine BEL (photographer)</creator>
<creator>Datta SHINDE (research participant)
Malavika TALUKDAR (video (photographer)
Bernard BEL (video (photographer)
Guy POITEVIN (annotator)</creator>
<creator>Gajarabai DAREKAR (transcriber)</creator>
<creator>Rajani KHALADKA (transcriber)</creator>
<date>2010-11-02</date>
<corpus type>
<format>AIFF</format>
<coverage>Box: name-Maharashtra</coverage>
<coverage>TGN: 1001895</coverage>
<rights>CRDO licence; rightsHolder = Bernard BEL</rights>
<rights>license:URI http://crdo.fr/licenceCRDO.php?version=1</rights>
</DocDC>
<DocMeta>
<authenticite>oui</authenticite>
<dureeConservation>P10000Y</dureeConservation>
<identifiantDocProducteur>crdo000759</identifiantDocProducteur>
<docRelation>
<typeRelation>version</typeRelation>
<sourceRelation>PAC</sourceRelation>
<identifiantSourceRelation>17714</identifiantSourceRelation>
</docRelation>
<docRelation>
<typeRelation>version</typeRelation>
<sourceRelation>Produc</sourceRelation>
<identifiantSourceRelation>crdo000759</identifiantSourceRelation>
</docRelation>
<docRelation>
<typeRelation>version</typeRelation>
<sourceRelation>CRDO-Aix</sourceRelation>
<structureDocument>Index.xhtml</structureDocument>
</version>
</DocMeta>
File description in the SIP

- <FichMeta>
  <encodage>PCM</encodage>
  <formatFichier>AIFF</formatFichier>
  <nomFichier>6.aif</nomFichier>
  <empreinteOri type="MD5">e06f2782f2992b15336a5734ba0d39c7</empreinteOri>
</FichMeta>

- <FichMeta>
  <encodage>PCM</encodage>
  <formatFichier>AIFF</formatFichier>
  <nomFichier>7.aif</nomFichier>
  <empreinteOri type="MD5">01503f8e495885a1ac18207bcd2a7d8</empreinteOri>
</FichMeta>

- <FichMeta>
  <encodage>PCM</encodage>
  <formatFichier>AIFF</formatFichier>
  <nomFichier>8.aif</nomFichier>
  <empreinteOri type="MD5">5dff69adca8fe53335af284b5860f9b</empreinteOri>
</FichMeta>

- <FichMeta>
  <encodage>PCM</encodage>
  <formatFichier>AIFF</formatFichier>
  <nomFichier>9.aif</nomFichier>
  <empreinteOri type="MD5">5cc33a222833db40c193e5d65135d040</empreinteOri>
</FichMeta>

- <FichMeta>
  <formatFichier>PDF</formatFichier>
  <nomFichier>BhaktiFAFaitForRehabilitation.pdf</nomFichier>
  <empreinteOri type="MD5">6c5814b22e48c3cc87a5db35cebe991f</empreinteOri>
</FichMeta>

- <FichMeta>
  <formatFichier>XML</formatFichier>
  <nomFichier>DESC/confidentiel.xml</nomFichier>
  <empreinteOri type="MD5">ac69593a1922738218ff7365c6e5a6d</empreinteOri>
</FichMeta>

- <FichMeta>
  <formatFichier>XML</formatFichier>
  <nomFichier>DESC/groupes.xml</nomFichier>
  <empreinteOri type="MD5">ebe5c11589dc2e708f245bf4163b6c17</empreinteOri>
</FichMeta>

- <FichMeta>
  <formatFichier>XML</formatFichier>
  <nomFichier>DESC/olac.xml</nomFichier>
  <empreinteOri type="MD5">3ff9bb6ca93d1e50884decba5217b2e9</empreinteOri>
</FichMeta>
Integrated access rights management
Integrated management of access rights

1. France takes advantage from a significant advance on archive law, owing to its legal framework establishing the status of ‘public archive’ and a uniform set of rules regulating the access to archived documents: *Code du patrimoine* (the Heritage Code), act of 15 July 2008, articles L213 1-5.

2. This framework implies a radical change of practice with respect to the long-term preservation and sharing of digital data, as any public archive shall be immediately in open access with the exception of 24 derogations applicable to certain categories of documents.

3. Most ethnographic and oral/linguistic material may be covered by a derogation (AR048) dealing with the protection of private information. In this case, access will be restricted to documents for which specific free consent forms have been signed by informants.

4. Art. L213-5: “*Any administration [...] is compelled to give reasons for denying access to a public archive.*”
25 years from the date of death of the individual. Documents disclosure of which would undermine medical confidentiality. (Code du Patrimoine, art. L. 213-2, I, 2)

25 years from the date of death of the person, whichever is shorter than the normal limit of 75 years. Marital status (birth or marriage). (Code du Patrimoine, art. L. 213-2, I, 4, e)

25 years from the date of death of the person, whichever is shorter than the normal period of 75 years. Minutes and inventories of public and ministerial officials. (Code du Patrimoine, art. L. 213-2, I, 4, d)

25 years from the date of death of the person, whichever is shorter than the normal period of 75 years. Documents relating to investigations conducted by the staff of the judicial police. (Code du Patrimoine, art. L. 213-2, I, 4, b)

25 years from the date of death of the person, whichever is shorter than the normal period of 75 years. Documents relating to cases before the courts, subject to special provisions relating to judgments, and execution of court decisions. (Code du Patrimoine, art. L. 213-2, I, 4, c)

50 years. Documents disclosure of which undermines the protection of privacy or for appreciation or value judgments about a person named or easily identifiable, or which reveal the behavior of a person under circumstances which might bring him/her injury. (Code du Patrimoine, art. L. 213-2, I, 3)
In this corpus, access to source video files is restricted to research scholars, teachers and students identified on the site and agreeing with the SLDR non-commercial licence. This restriction is legally justified by derogation code AR048 dealing with private life.

Public access to edited video will be envisaged after collecting the permissions of all participants in the documentation.

Making a digital document ‘public’ does not imply that it should be directly downloadable. This status is compatible with a non-commercial transaction granting access to any person registered as a ‘user’.
Participants agree with unrestricted access to low-resolution sound files (AAC) but their sources (WAV) are reserved to registered users approving the SLDR non-commercial licence.

Example of file sharing in high and low resolutions.
Access rights to documents are displayed in the descriptive metadata of an interoperable system.
Following up resource usage: users’ communities

This page displays the downloadings of an item and profiles of its users.

Access to this list is restricted to persons having produced items on SLDR or the ones who downloaded this particular item.