

Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France

Michel Jacobson, Olivier Baude

▶ **To cite this version:**

Michel Jacobson, Olivier Baude. Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France. Revue TAL : traitement automatique des langues, 2011, Ressources libres, 52, pp.47-69. halshs-01165884

HAL Id: halshs-01165884
https://shs.hal.science/halshs-01165884

Submitted on 20 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales[1] sur le français et les langues de France

Michel Jacobson * — Oliver Baude,*****

** Service interministériel des archives de France*
56, rue des Francs-Bourgeois
75003 Paris
michel.jacobson@culture.gouv.fr

*** Laboratoire Ligérien de Linguistique*
Université d'Orléans
olivier.baude@univ-orleans.fr

**** Délégation générale à la langue française et aux langues de France*

RÉSUMÉ. *Le programme «* Corpus de la parole *» est un projet en collaboration entre le ministère de la Culture et de la Communication et le CNRS qui vise à constituer une collection de ressources orales sur le français et les langues de France. Un portail Web offre un accès éditorialisé à cette collection. Cet article présentera les points principaux de l'organisation de ce programme, de la collecte des corpus aux aspects de pérennisation en passant par l'accès et la diffusion des données numériques.*

ABSTRACT. *"Corpus de la parole" is a collaborative project between the Ministry of Culture of France and the CNRS, which aims to build a collection of resources on French and other languages of France. A Web site provides an editorialised access to this collection. This article presents the main points of the organization of this program: the data collection, the access, dissemination and sustainability aspects of the digital data.*

MOTS-CLÉS : Corpus de la parole*, oral, archives ouvertes, OAIS.*

KEYWORDS: *Oral corpora, open archives, OAIS.*

1. Par commodité nous utiliserons le terme « orales » pour désigner des ressources orales mais aussi multimodales.

1. Contexte

L'origine du programme *Corpus de la parole* provient d'une volonté de l'Observatoire des pratiques linguistiques de la DGLFLF[2] de renouveler les initiatives de conservation et de diffusion des archives orales des linguistes commencées en 1911 avec la naissance des Archives de la parole de Ferdinand Brunot. Une centaine d'années plus tard, les questions de conservation et de diffusion relèvent également des aspects scientifiques de la constitution d'objets et de savoirs par une communauté de chercheurs.

Créé en 1999 au sein de la Délégation générale à la langue française, l'Observatoire des pratiques linguistiques a pour objectif de recenser, de développer et de rendre disponibles les savoirs relatifs à la situation linguistique en France, afin notamment de fournir des éléments d'information utiles à l'élaboration des politiques culturelles, éducatives ou sociales. Il a également pour but de faire mieux connaître un patrimoine linguistique commun constitué par l'ensemble des langues et des variétés linguistiques parlées en France, qui concourent à la diversité culturelle de notre pays. Doté d'un comité scientifique composé de linguistes, il soutient des projets ou des programmes de recherche dans le cadre d'appels à propositions thématiques ou de partenariats avec le CNRS ou les universités.

Le champ de l'observation est celui de la sociolinguistique et concerne les pratiques actuelles, qu'il s'agisse du français ou des autres langues parlées sur le territoire national : langues régionales ou langues issues des différentes vagues de migration. Depuis 2004, un des axes majeurs de l'activité de l'Observatoire est le développement, dans le cadre d'un partenariat Culture-CNRS, du programme *Corpus de la parole* qui a pour objectif la numérisation et la valorisation des corpus oraux (collections ordonnées d'enregistrements de productions linguistiques orales et multi-modales réalisées par des chercheurs) afin de permettre leur conservation et leur transformation en de véritables ressources linguistiques numériques, pour la recherche en sciences humaines, l'enseignement et l'ingénierie des langues. Il a permis, de 2006 à 2009, dans le cadre notamment du plan de numérisation du ministère de la Culture et de la Communication piloté par l'ex-MRT[3], de constituer et de numériser une collection de corpus oraux en français et en langues de France, mise à la disposition du public sur le site Internet *Corpus de la parole*[4], ouvert en février 2008.

Un *Guide des bonnes pratiques* (Baude, 2006) destiné aux chercheurs, a également été réalisé.

La démarche qui a présidé à la rédaction de guide était expressément centrée sur les pratiques des chercheurs afin de permettre une prise en compte des contraintes

2. Délégation générale à la langue française et aux langues de France.
3. Mission recherche et technologie du ministère de la Culture et de la Communication.
4. http://corpusdelaparole.in2p3.fr

scientifiques : méthodologie de collecte, opérations d'annotations et impacts sur les données.

Ce programme doit permettre, non seulement le développement d'une base de données patrimoniales sur l'oral, mais aussi le développement d'outils de traitement automatique des langues et d'ingénierie linguistique rendant possible l'interopérabilité des bases de données de grands corpus.

La DGLFLF, entre 2006 et 2008, a soumis une proposition à la MRT dans le cadre du « plan numérisation » du ministère de la Culture et de la Communication et a ainsi pu bénéficier d'une aide financière. Cette aide a permis la définition d'une organisation, de méthodes, la construction d'un prototype et son alimentation initiale en données ainsi que la numérisation de milliers de documents en différentes langues. La proposition de la DGLFLF était la constitution d'une base de données sur les langues de France constituée de ressources d'enregistrements audio ou vidéo récoltés au fil du temps par des scientifiques (linguistes, anthropologues...) pour leurs propres études. Le dossier soumis et accepté, montre une collaboration étroite entre la DGLFLF et le CNRS, ce dernier à travers deux acteurs principaux : le premier assurant un travail scientifique, le deuxième assurant un rôle de support technique.

La composante scientifique du projet est représentée par les deux fédérations de linguistique TUL[5] et l'ILF[6]. Son rôle consiste en l'identification des corpus existants et la constitution de nouveaux corpus sur des langues de France émanant de la communauté de recherche en linguistique. Ces corpus, dans la définition de départ du ministère de la Culture et de la Communication devant être composés de ressources existantes, anciennes, en danger et pour lesquelles la numérisation pouvait représenter une solution de sauvegarde, mais aussi de nouveaux corpus intégrant dès leur conception des nouvelles pratiques orientées vers la diffusion et la conservation. Pour ce projet, les fédérations de recherche en linguistique du CNRS exploitent leurs réseaux de contacts (laboratoires, projets, chercheurs) afin de faire émerger des propositions de collaboration, évaluent l'importance scientifique de ces propositions de numérisation avec l'aide d'un conseil scientifique dédié à ce programme et assurent la gestion et le suivi des candidatures retenues.

L'organisation du support technique du projet a été définie au départ en 2006 par la Direction de l'information scientifique (DIS) du CNRS. Celle-ci a divisé les tâches à réaliser en deux lots, et a confié chaque lot à des structures distinctes : l'INIST[7] a été chargée de la conception d'un portail d'accès et de valorisation des ressources constituées dans ce cadre et le CRDO-Paris[8] a été chargé de la gestion de ces ressources au sein de son entrepôt de ressources. Une architecture fonctionnelle

5. Typologie et universaux linguistiques.
6. Institut de linguistique Française.
7. Institut national de l'information scientifique et technique.
8. Centre de ressources pour la description de l'oral.

fondée sur le modèle des archives ouvertes[9] devait permettre la communication entre les deux modules (portail et entrepôt). Le travail du CRDO-Paris a été principalement un travail de définition d'une collection sur les langues de France puis de contrôle des ressources entrant dans la collection au fur et à mesure de leur collecte. Le reste des tâches utiles (gestion du catalogue, exposition des métadonnées, stockage, conservation, gestion d'accès, etc.) étant des missions directement inscrites dans la définition du CRDO.

2. *Les langues de France*

La base de données qui prendra finalement le nom de *Corpus de la parole* en écho aux « Archives de la parole » de Ferdinand Brunot[10], comporte des enregistrements de parole en français et en « langues de France[11] ». Il existe plus de 75 langues relevant de l'appellation « langues de France », les langues régionales (alsacien, breton...), les langues d'outre-mer (arawak, futunien...), les langues non territoriales (berbère, judéo-espagnol...) et la langue des signes française. Au fur et à mesure du temps la collection s'est enrichie par des langues entretenant un rapport étroit avec ces premières langues et qui faisaient l'objet de recherches au sein des laboratoires de linguistique (français parlé hors du territoire national, langues des signes émergentes...).

Tous les enregistrements de la collection sont issus de corpus récoltés par des linguistes dans le cadre de projets scientifiques.

La plus grande partie des enregistrements est audio, mais pour certaines langues (les langues des signes en particulier) ou pour certains usages (étude des interactions, étude du langage enfantin) la vidéo a également été utilisée par les chercheurs.

Les plus anciens enregistrements audio ont été faits avec des techniques analogiques sur des supports de type bandes magnétiques, puis, par la suite, sur des cassettes magnétiques. Les enregistrements les plus récents ont été faits avec des techniques numériques sur des supports de type disques optiques, disques magnéto-optiques, disques durs, mémoires flash. Parmi ces enregistrements, priorité a été donnée aux supports analogiques en fin de vie qui, en raison de leur ancienneté et de leur fragilité, étaient généralement les plus en danger. Pour autant, certains supports numériques ont des durées de vie très courtes, souvent inférieures aux supports analogiques et l'obsolescence technologique y est aussi plus présente et plus fréquente, ce qui a pu conduire parfois à récupérer en urgence le contenu de ces supports.

9. *Open Archives Initiative* (OAI).

10. Le centenaire a été marqué en 2011 à la BnF par une journée d'étude, le 17 juin 2011.

11. Les langues de France sont les langues régionales ou minoritaires parlées traditionnellement par des citoyens français sur le territoire de la République, et qui ne sont la langue officielle d'aucun État.

2.1. *Les projets existants similaires ou en relation*

Il existe bien sûr d'autres initiatives qui recoupent au moins en partie les préoccupations du projet *Corpus de la parole*. À une échelle locale on peut, bien sûr, citer des projets de laboratoires de recherche ou des projets trans-laboratoires (financés en particulier par l'ANR). Nous mettons dans ces catégories des projets sur le français tels que PFC (Phonologie du français contemporain), CFPP2000 (Corpus de français parlé parisien des années 2000), CRFP (Corpus de référence du français parlé), etc[12]. À une échelle plus large, il existe aussi des projets nationaux et internationaux portés par des institutions ou des communautés. Par exemple le projet CHILDES (*Child Language Data Exchange System*) est un projet créé dans le but de faciliter les échanges de corpus oraux au sein de la communauté qui étudie l'acquisition du langage par les enfants. Ce projet, pionnier à son époque, a défini des conventions pour la transcription, des outils pour la création et l'interrogation de corpus ainsi qu'une organisation du partage avec une banque de données des corpus partagés. Ce projet comporte aussi des données sur le français et les langues de France. D'autres projets internationaux plus récents ont vu le jour suite à une prise de conscience de l'aspect « en danger » de certaines langues ou de la difficulté de partage et de conservation du patrimoine scientifique et culturel que représentent ces données. Dans les projets sur les langues en danger nous pouvons citer par exemple le projet HRELP[13], le programme de l'UNESCO, le programme Sorosoro[14] de la fondation Chirac. Dans la mesure où il existe de nombreuses langues de France qui ne sont pas écrites ni enseignées, notamment à l'outre-mer en Nouvelle-Calédonie ou en Guyane, les interactions entre ces projets et le projet *Corpus de la parole* peuvent être étroites et les chercheurs qui travaillent sur ces langues participent parfois aussi à ces projets. Enfin, dans les projets institutionnels qui défendent des objectifs patrimoniaux, scientifiques et/ou culturels, nous pouvons citer les projets européens CLARIN (qui vise à la mise en place d'une infrastructure sur les ressources et technologies de la langue) ou FLaReNet (*Fostering Language Resources Network*). Le réseau public des archives en France collecte lui aussi des ressources linguistiques et cette communauté, en collaboration avec les autres institutions équivalentes dans les autres pays, participe également à des projets européens, en particulier pour favoriser l'accès à ces contenus (projets APEnet ou Europeana).

En résumé il existe de nombreux projets et organisations qui, pour des raisons diverses, financent et constituent des bases de données sur des ressources proches de celles collectées dans le cadre de *Corpus de la parole*. L'interaction avec ces projets et organisations se fait en général à deux niveaux :

---

12. Pour une liste plus complète des corpus existants on se reportera utilement à l'inventaire effectué en 2005 par Paul Cappeau et Magali Seijido « Les corpus oraux en français » (http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Presentation_Inventaire.pdf).
13. « *Hans Rausing Endangered Languages Project* » est un projet de la SOAS (*School of Oriental and African Studies*) de l'université de Londre.
14. Sorosoro signifie « souffle, parole, langue » en araki (langue du Vanuatu).

- au niveau du chercheur qui contribue en apportant au programme des ressources nouvelles, mais qui, pour d'autres raisons, participe aussi à d'autres projets aux objectifs variés ;
- entre les organisations ce sont plutôt les méthodes et les bonnes pratiques, qui font l'objet d'un partage et d'une discussion. Ainsi, depuis le départ du projet *Corpus de la parole*, les choix d'organisation, de formats, de normes de qualités ont été discutés avec des institutions (BnF, Archives de France, TGE-Adonis), qui pour des raisons extérieures au projet lui-même, participent à des groupes de standardisation et de normalisation et sont également présentes dans d'autres projets du même type.

2.2 *Les projets d'archivage et les théories*

Les développements parfois parallèles, parfois sécants, de la sociolinguistique et de la linguistique de corpus à partir des années soixante ont renouvelé les questions méthodologiques et théoriques de constitution et d'exploitation des données orales en linguistique. Cet article n'est pas le lieu d'une présentation de l'ensemble des points théoriques abordés ces cinquante dernières années, mais il convient de noter qu'il reste un clivage important entre un courant linguistique scindant les données et l'analyse des données, et un courant intégrant l'analyse aux conditions mêmes de productions des données. Selon ce second courant, les questions rencontrées lors des étapes de constitution, de conservation et de diffusion relèvent de positions théoriques. Ainsi, les choix de catalogage et de codage, les niveaux d'annotations et la description des éléments constitutifs d'un corpus sont autant de partis pris de théories trop souvent non explicites.

L'explicitation de la démarche de constitution et de ressources devient un exercice nécessitant une réflexivité placée au cœur même du travail scientifique. Les paragraphes suivants décrivent les lieux possibles et nécessaires de cette réflexivité et de cette explicitation.

3. Les préconisations

Afin d'effectuer le travail de préparation des ressources par leurs détenteurs, un certain nombre de préconisations ont été définies, portant tant sur les moyens à mettre en œuvre, que sur les résultats attendus.

3.1. *Les préconisations pour la numérisation des enregistrements*

Pour la numérisation des anciens supports analogiques, le CRDO-Paris a défini des critères de qualité minimaux. Ces critères, inspirés de ceux préconisés par

IASA[15] (IASA 2009) ont été validés par le conseil scientifique du programme *Corpus de la parole* en accord avec le département des archives sonores de la BnF et communiqués, *via* les fédérations de linguistique, aux chercheurs et laboratoires qui pratiquaient eux-mêmes la numérisation. Il s'agissait d'une préconisation contractuelle qui a donné lieu à l'élaboration d'une annexe technique, systématiquement présente dans les conventions de la DGLFLF. Le CRDO-Paris, qui pilotait également une partie des numérisations pour le compte des chercheurs et laboratoires qui le souhaitaient, appliquait aussi obligatoirement ces préconisations lors des opérations de numérisation à l'aide des équipements d'un laboratoire qui s'en était doté pour ses besoins propres (le LACITO[16]). Pour les enregistrements audio ces préconisations étaient les suivantes : échantillonnage 44,1 kHz au minimum (96 kHz au LACITO) ; quantification : 16 bits au minimum (24 bits au LACITO) ; copie droite sans retouche ; format WAV ; encodage : PCM[17].

3.2. *Les préconisations pour l'écriture des transcriptions*

Pour les annotations pouvant accompagner les enregistrements[18], le CRDO-Paris a défini, toujours après validation du conseil scientifique, des recommandations allant jusqu'à un modèle cible en XML. Ce modèle, exprimé dans une DTD XML, définit une structure minimale permettant :

- de coder la transcription d'un enregistrement ;
- d'ajouter une traduction en français (ce qui été demandé pour les langues autres que le français ou pour des transcriptions non orthographiques du français) ;
- de découper la transcription en segments (phrase ou groupe de souffle) ;
- de noter les repères temporels de début et de fin des segments.

Quelques raffinements du modèle permettent également d'indiquer le locuteur (utile pour les dialogues), le type de transcription (orthographique, phonétique, phonologique).

15. *International Association of Sound and Audiovisual Archives*.
16. Laboratoire de langues et civilisations à tradition orale.
17. Pulse-code modulation. Il s'agit d'un codage sans compression.
18. Le programme de la DGLFLF prévoit également une phase de valorisation des enregistrements à l'aide d'une ou plusieurs couches d'annotations (transcription, glose, traduction, annotations morphosyntaxiques, syntaxiques ou autres).

```
<S id="ESTAQUEs1" who="INT">
    <FORM kindOf="ortho">Se akodra de antes de la gera, kómo, kómo era la vida en el Estaque ?
        Ké aziya la djente...</FORM>
    <TRANSL xml:lang="fr">Vous vous rappelez, avant la guerre, comment, comment était la vie à
        l'Estaque ? Que faisaient les gens...</TRANSL>
    <AUDIO start="0.0000" end="9.5063"/>
</S>
<S id="ESTAQUEs2" who="LOC">
    <FORM kindOf="ortho">En Estaque bivimo bivimos komo los... la djente ke viviyan en el
        Estaque,</FORM>
    <TRANSL xml:lang="fr">A l'Estaque nous avons vécu comme les... les gens qui vivaient à
        l'Estaque,</TRANSL>
    <AUDIO start="9.5063" end="17.9336"/>
</S>
```

Figure 1. *Transcription extraite du corpus* « Judeo-Spanish in France Archive »

Ce modèle minimal peut être atteint soit directement, soit en passant par des formats et outils qui permettent de faire une annotation plus riche et plus fine. En particulier, les formats en sortie des outils Transcriber[19] et ITE[20] peuvent être directement exploités dans le cadre du projet, la transformation vers le format cible étant alors complètement automatisée. D'autres formats, tels que ceux utilisés par les outils CLAN ou ELAN, doivent faire l'objet d'une normalisation afin d'être transformés de manière souvent *ad hoc* dans le format cible. Dans ce dernier cas, les deux formats sont conservés : le format d'origine et le format cible.

Ces recommandations ne portent que sur la forme à utiliser pour exprimer les transcriptions. Aucune indication ni directive n'est donnée pour expliquer aux chercheurs comment ils doivent transcrire leurs enregistrements et en donner une traduction en français. Les linguistes ont parfois établi des conventions accompagnées de manuels[21], mais d'une langue à l'autre (du français à la langue des signes française, par exemple), ou d'un domaine linguistique à un autre (de la phonétique à la dialectologie, par exemple), ces conventions sont peu partagées. Il est en revanche conseillé d'identifier, sous forme d'une référence dans les métadonnées, la ou les ressources qui décrivent de manière explicite l'ensemble des conventions utilisées.

3.3. *Les consignes pour la description des ressources*

La description des ressources (les enregistrements et les transcriptions) a également donné lieu à l'élaboration de préconisations. Cette description repose sur un jeu de métadonnées qui doivent suivre le schéma XML défini par OLAC[22]. Ce

19. Transcriber (http://trans.sourceforge.net/).
20. ITE Interlinear Text Editor (http://michel.jacobson.free.fr/ITE/).
21. Conventions pour ESLO (http://eslo.in2p3.fr/).
22. *Open Language Archives Comunity*.

schéma reprend ceux du Dublin-core et du Dublin-core qualifié[23] auxquels sont ajoutés cinq attributs associés à des vocabulaires contrôlés (*role*, *language*, *linguistic-field*, *linguistic-type* et *discourse-type*). Les recommandations précisent la manière d'utiliser ce schéma OLAC pour décrire les ressources (les éléments obligatoires, facultatifs, des explications, des exemples, etc.).

4. Organisation de la production

La première opération consiste pour les fédérations de linguistique à identifier dans leurs réseaux de contact, ou par l'intermédiaire d'un appel à projets, les corpus existants qui pourraient entrer dans le cadre du projet. Puis elles prennent contact avec les responsables de ces corpus afin d'évaluer l'intérêt scientifique et la charge de travail (en temps, ressources humaines, budget) que représente l'entrée du corpus (tout ou partie) dans la collection. C'est aussi le moment où une expertise des aspects juridiques est nécessaire. Il s'agit tout d'abord d'évaluer si les ressources identifiées sont susceptibles de poser des problèmes en termes de propriété intellectuelle et de respect de la vie privé. Cette expertise nécessite la prise en compte d'informations précises qui ne peuvent être formulées que dans le cadre de l'explicitation de la démarche suivie par les chercheurs tout au cours de leur projet de recherche. Ainsi, la description de la méthodologie de collecte et de traitements des données ne peut se faire sans une présentation précise des cadres théoriques mobilisés, par exemple dans les opérations de catégorisation des participants et de leurs productions langagières.

En fonction de l'évaluation des risques, des mesures peuvent être envisagées telles que la recherche des autorisations ou l'anonymisation. Si aucune solution ne permet la libre accessibilité à la ressource, celle-ci ne pourra tout simplement pas entrer dans le cadre strict du projet. Cette contrainte juridique est en effet dépendante des objectifs du « plan numérisation » du ministère de la Culture et de la Communication, qui conditionne le financement de la numérisation à la mise à disposition des données publiques. Ceci n'interdit nullement d'archiver les ressources en question dans le respect des conditions d'accessibilité définies par le Code du patrimoine[24], mais elles ne pourront alimenter le réservoir C*orpus de la parole* qu'à l'issue d'une période plus ou moins longue. Si le corpus en question représente bien un intérêt, un accord précise que les responsables du corpus cèdent de manière non exclusive les droits de représentation et de reproduction des ressources qui seront numérisées dans ce cadre à la DGLFLF afin que cette dernière puisse alimenter le portail avec ces ressources.

23. Dublin-core ou norme ISO 15836.
24. « Que sont les archives ? » dans la lettre d'information de l'InSHS n°13 de septembre 2011 (http://www.cnrs.fr/inshs/Lettres-information-INSHS/lettre_infoINSHS_13.pdf).

4.1. *Numérisation*

Le travail de numérisation est effectué, suivant les cas, soit directement par le chercheur s'il dispose des outils et des compétences nécessaires, soit par un tiers qui souvent est le laboratoire du chercheur. Les consignes précisées plus haut sont transmises aux chercheurs et le CRDO apporte son expertise et un accompagnement à toutes les étapes du projet.

4.2. *Collecte*

La collecte des ressources est une tâche assurée par les fédérations de linguistique. Elle consiste en un suivi de planning pour les différents contributeurs avec des échanges et des relances jusqu'à obtenir la livraison complète des enregistrements numérisés correctement documentés et éventuellement, suivant les cas, accompagnés de fichiers de transcriptions et traductions. À partir de 2009, les fédérations ont recruté sur contrat un ingénieur pour assurer l'accompagnement des projets auprès des chercheurs et des laboratoires. Une fois cette collecte faite, les fédérations de linguistique effectuent une livraison du corpus au CRDO-Paris. La livraison doit respecter les préconisations techniques y compris le respect des consignes de nommage et de formats des fichiers. Elle s'effectue comme un simple dépôt de fichiers sur le serveur dans une zone réservée à cet effet.

La tâche d'accompagnement été particulièrement soignée dans le but de respecter l'ensemble des contraintes relevant de choix méthodologiques et théoriques des chercheurs. Il ressort de cette expérience le constat d'une très grande hétérogénéité des pratiques qui demande à être respectée afin de préserver la chaîne des opérations qui lie la collecte à la diffusion en passant par l'analyse et l'exploitation. Les problèmes rencontrés ont été systématiquement étudiés au sein du conseil scientifique.

4.3. *Contrôle*

Le contrôle de la livraison est effectué à réception des ressources par le CRDO-Paris. Une première analyse qualité est effectuée pour s'assurer de la complétude des informations : à chaque fichier livré doit correspondre un jeu de métadonnées et chaque jeu de métadonnées doit décrire un fichier unique et distinct. La présence d'informations obligatoires dans les métadonnées est également vérifiée : nom du déposant, identification de la langue... Enfin le CRDO-Paris contrôle la bonne formation des fichiers (enregistrements, annotations et métadonnées) et enrichit les métadonnées avec des informations de nature technique. Suivant la nature du contrôle ce dernier est effectué par un documentaliste ou par un informaticien.

La vérification des fichiers d'enregistrement est faite à l'aide d'un outil qui extrait les informations de formatage (format de l'enveloppe, codage des contenus,

fréquence d'échantillonnage, taille des échantillons, nombre de canaux). Ce programme émet des alertes lorsque les critères définis ne sont pas respectés, si des plages silencieuses dépassent une durée seuil[25] ou encore si des informations inattendues sont présentes dans le document (par exemple : des métadonnées, des jalons temporels).

La vérification des fichiers de métadonnées et des fichiers d'annotations passe systématiquement par des technologies de validation de schémas XML.

Parallèlement à ces contrôles, essentiellement techniques, une autre vérification est effectuée par un opérateur humain de formation documentaliste, dont le rôle est de vérifier la cohérence des métadonnées, leur complétude et leur exactitude et d'entretenir un catalogue de ressources homogène et compréhensible. Au besoin les métadonnées seront donc complétées et normalisées, éventuellement après des échanges auprès du déposant de la ressource pour récupérer l'information et la valider.

L'enrichissement automatique des métadonnées concerne principalement les informations techniques à propos du fichier lui-même (type mime, type DCMI, durée pour les enregistrements, liaison au schéma pour les annotations) ainsi que quelques informations de gestion telles que la date de dernière modification des métadonnées, le lien à la collection, les URLs d'accès.

4.4. *Versement*

Une fois la ressource contrôlée, elle est mise en ligne sur le site du CRDO-Paris avec une restriction d'accès qui n'autorise que les administrateurs et le déposant à la consulter. Le déposant est alors encouragé à vérifier si sa ressource et la description qui en est faite sont correctes avant de donner son accord pour publication. Une fois cet accord obtenu, la restriction d'accès est supprimée et la ressource est accessible par tous et notamment sur le portail *Corpus de la parole*.

5. Description de l'architecture technique

Depuis le début du projet, le portail d'accès aux ressources et l'entrepôt de ressources sont séparés logiquement et physiquement. Les liens entre ceux-ci se font à l'aide du protocole OAI-PMH[26] défini par les « archives ouvertes ».

La conception du portail a été confiée au départ du projet à l'INIST. Sur la base de choix validés par la DIS et par l'Observatoire des pratiques linguistiques, il a

25. La durée paramétrable a été fixé arbitrairement à 10 secondes et permet d'alerter le technicien afin qu'il puisse vérifier si ces plages silencieuses sont intentionnelles (anonymisation) ou accidentelles (mauvaise transmission, défaut à la numérisation, etc.).
26. *Open Archives Initiative - Protocol for Metadata Harvesting*.

développé un site fondé sur un gestionnaire de contenu largement répandu (SPIP[27]). Différents rédacteurs peuvent ainsi saisir des contenus avec éventuellement une politique de validation. L'INIST a développé et mis en œuvre une charte graphique pour le site et a été conduit à développer quelques scripts (écrit en langage PHP et utilisables sous forme de « plugin » dans SPIP) afin de permettre d'afficher dans les pages du site des moteurs de recherche et des animations multimédia (pour la consultation des ressources mélangeant audio, vidéo et annotations). Par la suite la charte graphique du site a été revue par l'atelier de création « des signes graphiques » (figure 2) et l'ensemble des moteurs de recherche et outils de consultation ont été réécrits par le CRDO-Paris. Enfin, le portail lui-même après avoir été hébergé les premières années sur le serveur du CRDO-Paris l'est aujourd'hui sur les serveurs du centre de calcul de l'IN2P3 au sein de la grille de services du TGE-Adonis.

Pour faciliter le travail et accroître l'indépendance du portail, une base de données a été conçue qui stocke les métadonnées moissonnées à l'aide du protocole OAI-PMH de sorte qu'une fois ce moissonnage effectué les moteurs de recherche du portail n'aient plus besoin d'avoir recours aux services du CRDO-Paris pour retrouver de l'information. La périodicité du moissonnage a été fixée à sept jours (après avoir été quotidienne dans les premiers temps) et la technique utilisée est différentielle, c'est-à-dire qu'on ne moissonne que les changements intervenus depuis la dernière moisson, plutôt que complète (ce qui était fait dans un premier temps quand les volumes à moissonner étaient plus petits).



Figure 2. *Page d'accueil du site* Corpus de la parole

27. Système de publication pour l'Internet.

5.1. *Les moteurs de recherche*

Trois moteurs de recherche ont été développés pour le portail. Les deux premiers permettent une recherche dans les métadonnées des ressources, le dernier dans les transcriptions elles-mêmes.

Le premier moteur de recherche (cf. figure 3) permet des requêtes assez classiques dans les métadonnées en utilisant comme critères de recherche les quinze catégories définies dans la norme Dublin-core (la langue en tant qu'objet d'étude de la ressource pouvant être surajoutée aux autres critères de recherche). Le résultat retourné par ce moteur de recherche est composé d'une liste de ressources faisant apparaître comme seuls critères : le titre, le nom du chercheur et la langue en tant qu'objet d'étude (cf. figure 3). Cette liste permettant elle-même de donner accès *via* des liens à l'ensemble des métadonnées ainsi qu'aux ressources elles-mêmes.

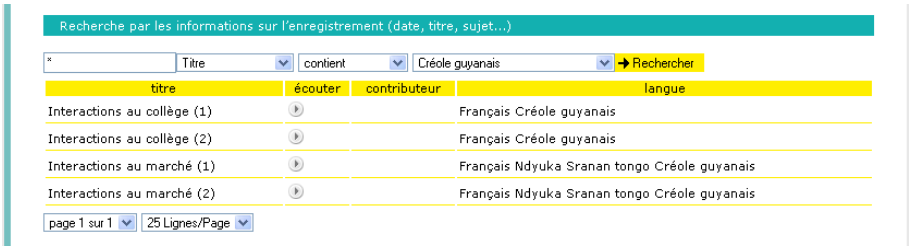


Figure 3. *Moteur de recherche dans les métadonnées*

Le deuxième moteur, mono-critère (la langue en tant qu'objet d'étude), retourne soit une liste de ressources comme le précédent moteur (cf. figure 3), soit une carte géographique sur laquelle les ressources qui sont géoréférencées sont indiquées (cf. figure 4). L'API Google Maps est utilisée pour l'affichage et la navigation dans la carte. De la même manière que pour le précédent moteur la liste tout comme la carte permettent de donner accès *via* des liens à l'ensemble des métadonnées ainsi qu'aux ressources elles-mêmes.

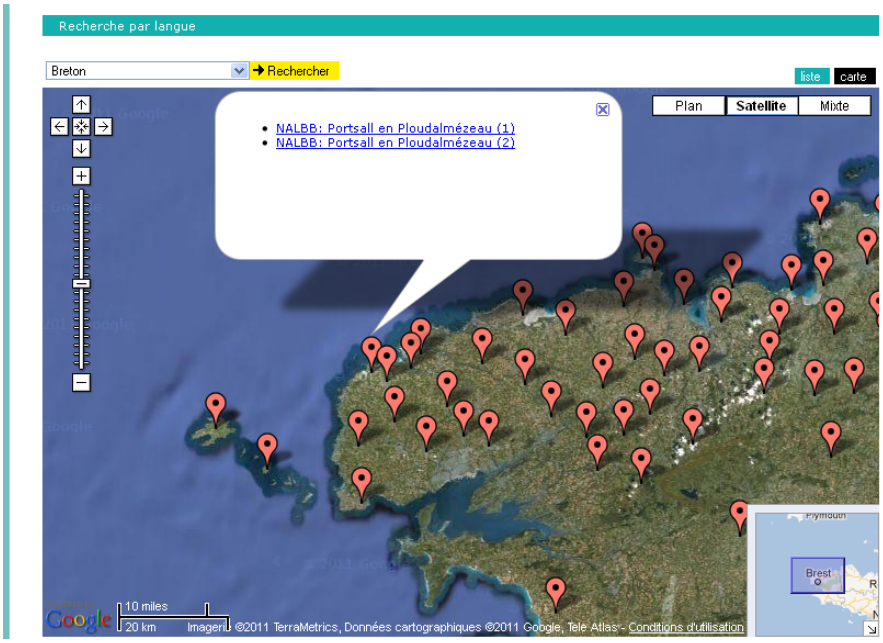


Figure 4. *Moteur de recherche géographique*

Le dernier moteur de recherche exploite cette fois l'ensemble des annotations et non plus les métadonnées, afin de pouvoir chercher un mot ou un motif soit dans la traduction soit dans la transcription. Le résultat retourné présente l'ensemble des segments qui contiennent ce mot ou ce motif (cf. figure 5).

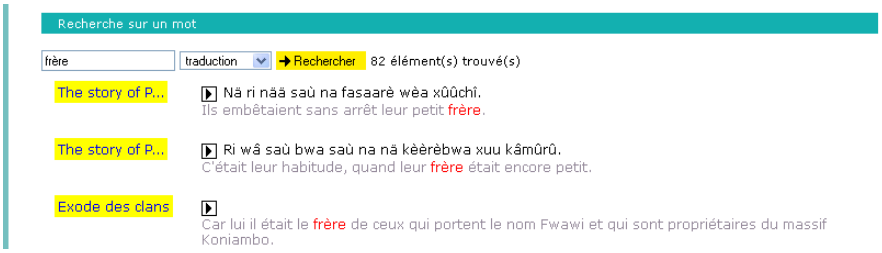


Figure 5. *Moteur de recherche dans les annotations*

5.2. *La consultation d'une ressource*

Lorsqu'une ressource est choisie par l'utilisateur, celui-ci peut la consulter par une interface multimédia (cf. figure 6). Dans tous les cas la page de consultation inclut un lecteur audiovisuel qui donne accès à l'enregistrement audio ou vidéo, ainsi qu'une fiche documentaire présentant une partie des métadonnées. Dans les cas où la ressource possède une transcription/traduction, celle-ci est également présentée dans la page et sa lecture est synchronisée avec le lecteur de sorte que cliquer sur le texte permet de démarrer la lecture de l'enregistrement à ce moment, et déplacer le curseur du lecteur positionne le texte à l'emplacement correspondant au moment choisi.

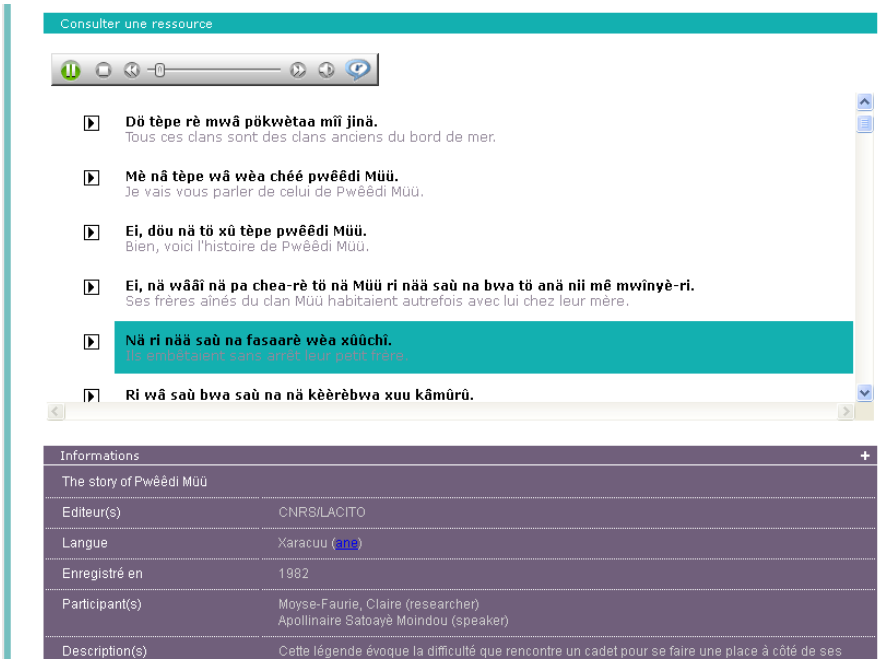


Figure 6. *Interface de consultation d'un document*

Les technologies qui ont été choisies pour réaliser ces fonctionnalités sont fondées sur le plugin Real, mais le nouveau HTML 5, qui offre la possibilité de coder directement en HTML des balises audio et vidéo, devrait permettre de s'affranchir de la technologie propriétaire de Real.

6. Prise en charge de l'archivage à long terme

Depuis mi-2008, le TGE-Adonis s'est engagé dans un programme d'archivage pérenne de données et documents numériques issus de la communauté des sciences humaines et sociales (SHS). L'organisation que le TGE-Adonis met en place dans le cadre de ce programme s'inspire du modèle de la norme OAIS[28]. Le modèle fonctionnel de cette norme distingue différentes briques illustrées en figure 7 et centrées autour :

- de l'entrée des archives : cette brique aborde le traitement des paquets d'informations versés par les producteurs d'archives. Elle comporte des mécanismes de préparation, de transmission, de contrôle, de rejet, de conversion de format, etc. Une fois le paquet d'informations validé et complété, cette brique le met à disposition des briques stockage et gestion de données ;
- de l'accès aux archives : cette brique traite des mécanismes d'accès, de consultation et de livraison des informations disponibles dans le système d'archivage (métadonnées et contenus). Elle comprend la mise à disposition d'un système de recherche dans les métadonnées, de sélection dans les résultats de la recherche, d'une interface de consultation et éventuellement un mécanisme de suivi des commandes jusqu'à leur livraison ;
- du stockage : cette brique traite de la conservation des informations à partir du moment où elles sont mises à sa disposition par la brique entrée et jusqu'à leur éventuelle destruction. C'est cette brique qui traite du choix des supports, de la gestion du contrôle de l'intégrité des données et de la gestion des migrations (rafraîchissement de supports, duplication et ré-empaquetage) ;
- de la gestion de données : cette brique assure la conservation, la mise à disposition et la mise à jour des informations descriptives (métadonnées) associées aux contenus d'informations conservés par la brique stockage ;
- de la planification de la pérennisation : cette brique assure une veille technologique et propose des recommandations, des évolutions et des stratégies pour prévenir l'obsolescence et garantir l'accès, sur le long terme, aux informations ;
- de l'administration : cette brique permet d'assurer l'exploitation de l'ensemble du système d'archivage électronique et traite en particulier de la gestion des utilisateurs au sens de leurs droits d'accès.

Le TGE-Adonis, sur la base d'une étude sur l'hébergement de services informatiques et de données numériques pour les SHS en France[29], a choisi d'adosser ce service à deux grands centres informatiques existants. Les centres choisis sont le

28. Modèle de référence pour un système ouvert d'archivage d'information (OAIS). Norme ISO 14721:2003

Centre informatique national de l'enseignement supérieur (CINES) et le centre de calcul de l'Institut national de physique nucléaire et de physique des particules (CC-IN2P3). Le CINES, qui avait déjà une expérience dans le domaine de la conservation avec les thèses et les données de numérisation des revues du portail Persée, s'est vu confier la brique d'entrée alors que le CC-IN2P3, qui avait déjà une forte expérience en hébergement et services, se voyait confier la brique d'accès correspondante. Les autres briques étant partagées entre eux avec une légère prédominance du CINES.

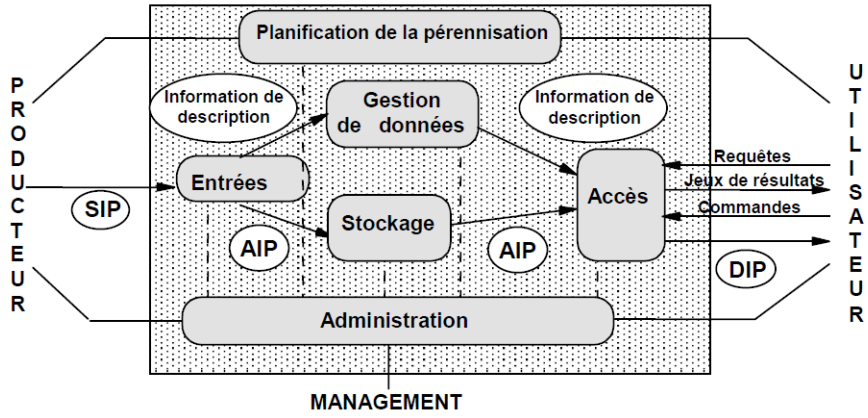


Figure 7. *Les fonctions du modèle OAIS*

Pour la mise en place de cette architecture, le TGE-Adonis a pris contact avec les Archives de France et a souhaité démarrer une première expérience sur les données orales gérées par le CRDO. Pour mener à bien cette expérience un groupe de travail a été constitué en rassemblant des membres des deux centres informatiques (CINES et CC-IN2P3), des membres des deux antennes du CRDO (Paris et Aix), un représentant des Archives de France, un chef de projet du TGE-Adonis et un consultant extérieur[30]. Après un peu plus d'un an de mise au point, l'architecture est entrée dans une phase de production réelle le 22 juin 2010 et offre un certain nombre de services au CRDO et donc au projet *Corpus de la parole*.

29. Étude Olof Bärring, « *Hosting of IT services and data for Human and Social sciences in France* », version 1.0, janvier 2008. Rapport final de l'étude commandée par le TGE Adonis au CERN. Document non public, communicable par le TGE Adonis sur demande.
30. Huc C, Habert B, Building together digital archives for research in social sciences and humanities, Social Science Information, vol. 49, n°3, septembre 2010, p. 415-443.

6.1. *Contrôle des entrées*

Dans l'organisation mise en place, les services servant (CRDO-Paris et CRDO-Aix) versent leurs archives au CINES *via* un protocole[31] qui prévoit dans leur ordre d'émission, les messages : bordereau de versement, accusé de réception, certificat d'archivage ou avis d'anomalie.

Entre le CRDO et le CINES, trois scénarios ont été définis : a) le transfert initial d'un objet, b) la modification de la description d'un objet et c) le transfert d'une nouvelle version d'un objet existant. Un dernier scénario dit de « restitution » a également été défini. Ce dernier ne doit intervenir qu'en cas de défaillance du CINES ou de fin de contrat entre le CNRS et le CINES afin de confier les archives dont le CINES a la responsabilité de la conservation à une autre organisation.

L'objet que l'on transfère une première fois et dont on peut par la suite modifier la description et/ou ajouter des versions, correspond à une unité d'archive dans laquelle peut se trouver un ensemble plus ou moins grand de fichiers. Chaque service versant doit donc définir de quoi est composé cet objet et le décrire dans un document[32] pour le CINES.

Les ressources du *Corpus de la parole* ne peuvent être que de l'un des trois types suivants :

- un enregistrement (audio ou vidéo) ;
- une annotation d'enregistrement (transcription, traduction, etc.). Cette annotation doit être un document structuré en XML et suivre le schéma défini dans les consignes du projet ou suivre l'une des deux DTD : celle du logiciel Transcriber ou du logiciel ITE ;
- une collection (rassemblement de plusieurs enregistrements et annotations).

L'identification du format de chaque fichier est une des informations techniques obligatoires à renseigner dans le bordereau de versement. En effet, le CINES ne peut engager sa responsabilité que sur des fichiers dont il connaît la structure et dont il peut s'assurer qu'ils respectent bien cette structure. Pour pouvoir prendre en charge les documents du CRDO, le CINES a été conduit à évaluer les formats et codages audiovisuels utilisés au CRDO et à faire évoluer sa plate-forme afin de pouvoir les ajouter à la liste des formats pris en charge. Cette étude[33] a identifié un certain nombre de formats de représentations acceptables tant pour l'audio que pour la vidéo. Par exemple pour l'audio, les formats acceptables par le CINES et utilisés dans *Corpus de la parole* sont le format WAV avec un encodage PCM ainsi que le

31. Le protocole s'inspire du « Standard d'échange de données pour l'archivage ».
32. *Project Preservation Description Information* (PPDI) terme de la terminologie OAIS.
33. « Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles ».

format FLAC[34]. Pour les vidéos, les formats conseillés par le CINES et que le programme *Corpus de la parole* a finalement retenus sont le format conteneur MKV avec le codec vidéo H264 et l'encodage FLAC pour l'audio ainsi que le format conteneur MPEG-4 avec le codec vidéo H264 et l'encodage AAC pour l'audio.

Tous les fichiers transmis au CINES, bien qu'ils aient déjà fait l'objet d'une validation par le CRDO-Paris avec ses propres outils, font l'objet à leur réception d'une deuxième validation par le CINES et des messages d'anomalies sont retournés s'ils ne suivent pas toutes les spécifications du format dont ils se réclament ou si les formats utilisés ne font pas partie de la liste des formats acceptables par le CINES.

Lors de la réception des fichiers le CINES effectue aussi un contrôle d'intégrité en comparant les empreintes des fichiers calculées par le service versant lors du transfert et renseignées dans le bordereau de versement avec les empreintes qu'il calcule lui-même sur les fichiers reçus. Ce contrôle est important afin de s'assurer que le fichier reçu correspond bien au fichier envoyé et qu'il n'y a pas eu par exemple d'erreur de communication lors du transfert.

À réception d'un paquet le CINES exécute deux opérations :

- il attribue au paquet un jeton d'horodatage calculé sur la base de source de temps extérieure et fiable. Ce jeton d'horodatage peut servir de garantie opposable d'antériorité en permettant de démontrer qu'une ressource existait bien à partir d'une date et d'une heure précises et certifiées ;
- il associe au paquet un identifiant pérenne de type ARK[35]. Cet identifiant est unique non seulement au sein de la plate-forme d'archivage du CINES mais également hors de ce contexte. Il permet ainsi de garantir une pérennité même à travers les différents organismes qui prendront par la suite en charge cette archive. C'est aussi un identifiant qui permet la « citabilité » de la ressource.

Les URL ARK sont composées de deux parties : la première partie appelée NMA[36] est une URL d'accès qui va permettre la localisation de la ressource. Cette partie est facultative, peut être modifiée (par exemple en cas de changement d'organisme en charge de cette gestion) ou multiple (gestion confiée à plusieurs organismes en même temps). La deuxième partie est l'identifiant à proprement parler de la ressource. Cet identifiant est obligatoire et ne peut ni changer ni être recyclé en cas de suppression de la ressource. Il se décompose en un identifiant de l'organisme qui affecte les identifiants aux ressources NAAN[37] suivi d'un identifiant de la

34. Ce format a été utilisé pour les enregistrements qui dépassent le volume accepté par le format WAV.
35. *Archival Resource Key*.
36. *Name Mapping Authority* ou autorité d'adressage.
37. *Name Assigning Authority Number* ou autorité d'assignement de noms. Par exemple, le NAAN du CINES est 87895, celui de la BNF est 12148. Une liste complète est maintenue à l'url http://www.cdlib.org/uc3/naan_registry.txt.

ressource, suivi éventuellement d'un « *qualifier* » permettant de gérer la granularité de la ressource, ses versions, ses formats, etc. Par exemple l'identifiant ARK attribué par le CINES pour un enregistrement en alsacien effectué à Aschbach dans le Bas-Rhin en 1980 dans le cadre des atlas linguistiques est ark:/87895/1.5-124712.

Dans l'état actuel du projet, l'autorité d'assignation n'étant pas encore définie, la citation d'une ressource se limite donc à sa seule identification (comme on le ferait avec le numéro ISBN d'un ouvrage) et non pas à sa localisation. Afin de permettre cette localisation le CRDO-Paris maintient des identifiants OAI. Ainsi la ressource du dernier exemple est localisable en interrogeant l'entrepôt du CRDO-Paris avec l'identifiant oai:crdo.vjf.cnrs.fr:crdo-HUD_0003_SOUND[38] par le protocole OAI-PMH.

6.2. *Transfert de responsabilité pour la conservation des informations*

Une fois tous les contrôles effectués par le CINES, si ce dernier juge acceptable le paquet d'informations versé, il retourne au service versant un certificat d'archivage afin de lui notifier le transfert de responsabilité et de lui communiquer l'identifiant permettant de récupérer au besoin l'archive. Le service versant (CRDO-Paris) peut alors se débarrasser des fichiers sur son système d'information puisqu'il n'a plus la responsabilité de leur conservation.

La responsabilité de la conservation par le CINES couvre toutes les tâches concourant à s'assurer que l'information transmise reste intègre et lisible à travers le temps, de s'assurer de son authenticité, de son accessibilité et éventuellement de sa confidentialité. Cela entraîne de nombreuses précautions à prendre en matière de stockage : l'information doit être copiée en plusieurs exemplaires, sur plusieurs sites distants, éventuellement sur des supports de différentes natures. Ces supports doivent être surveillés régulièrement et des migrations de supports devront être planifiées afin de lutter contre leur vieillissement et leur obsolescence. Des migrations de formats devront également être planifiées en cas d'obsolescence des formats d'origine.

De nombreuses autres tâches incombent au CINES du simple fait de prendre en charge la responsabilité de la conservation de ces informations. Afin de s'assurer que le CINES possède les compétences nécessaires en la matière et offre des garanties quant à leur mise en œuvre, il lui a été demandé, avant de pouvoir assurer cette mission, d'obtenir un agrément du ministère de la Culture et de la Communication. Cet agrément obtenu en décembre 2010 porte sur la conservation d'archives publiques numériques courantes et intermédiaires. Cet agrément permet en

38. La requête pour récupérer les métadonnées OLAC de la ressource en OAI-PHM est http://crdo.vjf.cnrs.fr/crdo_servlet/oai-pmh?verb=GetRecord&identifier=oai:crdo.vjf.cnrs.fr:crdo-HUD_0003_SOUND&metadataPrefix=olac

particulier de s'assurer que le service respecte bien les normes en usage dans la profession : à savoir la norme OAIS ainsi que la norme NF-42-013.

6.3. *Accès*

Une fois le versement accepté par le CINES, une copie est envoyée au CC-IN2P3 afin d'offrir un service d'accès plus large que celui que propose le CINES qui ne s'adresse qu'au seul service versant. Cela permet, par la même occasion, d'avoir une autre copie distante des archives (le CINES est situé à Montpellier alors que le CC-IN2P3 est situé à Villeurbanne près de Lyon).

Quand le CC-IN2P3 reçoit un paquet d'informations en provenance du CINES, celui-ci analyse son contenu et alimente son outil de gestion (Fedora-commons) en créant un nouvel objet. À cet objet sont associées les métadonnées fournies dans le paquet. Les liens de filiation (liens entre la transcription et l'enregistrement ou entre un enregistrement ou une transcription et une collection) sont reconstruits en rattachant les nouveaux objets aux objets parents précédemment archivés. Il en est de même pour les liens de version. Les relations inverses sont également ajoutées automatiquement. Ces relations sont exprimées dans l'outil Fedora-commons sous forme RDF avec des verbes du type isSonOf, isFatherOf, hasNextVersion, hasPreviousVersion, etc.

Chaque fichier (enregistrement ou transcription) est placé dans un système de fichiers et devient accessible par le Web *via* une URL. Pour certains types de fichiers, des formats de diffusion ont été définis de sorte que l'accès à une même information puisse se faire dans plusieurs formats correspondant à des usages différents. Pour le format d'archivage WAV/PCM nous avons ainsi défini au CRDO-Paris, trois formats de diffusion : 1) le format d'origine ; 2) une version dégradée en WAV/PCM à 22 kHz/16bits/mono ; 3) et une version dégradée au format MP3. Enfin une dernière version au format RealMedia devrait être prochainement disponible pour une diffusion par un serveur de streams.

Des mécanismes d'authentification associés à des mécanismes de restriction d'accès sont prévus afin de pouvoir assurer pendant le temps nécessaire la confidentialité des ressources qui le demandent. Ces mécanismes ne sont pas encore en production au moment de l'écriture de l'article, mais les ressources présentes sur le portail *Corpus de la parole* sont jusqu'à présent sans restriction d'accès. Au-delà même de cette liberté d'accès, les ressources du portail sont également toutes assorties d'une mention de licence Creative-commons[39] afin de préciser leurs conditions de réutilisation.

Une fois que l'ingestion des informations du paquet reçu dans Fedora-commons est terminée, l'objet devient accessible et le CRDO peut récupérer les informations

39. CreativeCommons est une organisation, qui un peu sur le modèle du mouvement des logiciels libre a défini des licences permettant la cession de certains droits aux utilisateurs pour des œuvres.

d'enrichissement par l'emploi des fonctions OAI-PHM de l'entrepôt OAI de Fedora-commons. En particulier, les informations de date d'archivage, d'identifiant ARK ainsi que les différentes URL d'accès aux fichiers sont récupérées pour enrichir et modifier les métadonnées présentes au CRDO-Paris et donc également sur le portail *Corpus de la parole*.

7. État du chantier

En 2012, le réservoir *Corpus de la parole* contient plus de mille heures d'enregistrements audio et plusieurs centaines de transcriptions ou traductions. Parmi les 78 langues de France, 42 sont présentes. Tous les documents présents sont accompagnés de leurs métadonnées et ils sont tous accessibles par le site portail selon les conditions présentées dans les paragraphes précédents.

Cette évaluation, purement quantitative, ne signifie rien sans la prise en compte des phases du projet consacrées à l'interopérabilité de données nativement hétérogènes. L'architecture mise en place a permis de tester les propositions de catalogage, de codage et d'archivage de corpus collectés et exploités par des communautés scientifiques. La diversité des laboratoires, des projets et des données assure la représentativité des pratiques actuelles des chercheurs.

8. Perspectives et conclusions

Cette expérience en cours devrait prochainement bénéficier d'une convention entre la BnF, la DGLFLF et le CNRS. La volonté d'une coopération entre la BnF et le CNRS n'est pas récente. Déjà en 1979, le CNRS et le département de la Phonothèque et de l'Audiovisuel de la Bibliothèque nationale avaient signé une convention. Celle-ci stipulait que les chercheurs œuvrant dans le cadre des Atlas linguistiques versaient des enregistrements et leur fiche descriptive à la BnF. La BnF en effectuait une copie destinée à la conservation et restituait aux chercheurs des bandes magnétiques vierges afin qu'ils puissent continuer leurs enquêtes. La convention précisait que les bandes magnétiques seraient consultables à la Bibliothèque nationale et dans une institution régionale afin qu'un public de chercheurs puisse avoir accès à un fonds patrimonial sonore normalisé et répertorié[40].

Les avancées technologiques et la modification des pratiques des chercheurs offrent aujourd'hui l'opportunité de construire une nouvelle convention qui s'appuierait sur l'expérience acquise dans le cadre du programme *Corpus de la parole*. L'architecture proposée permet notamment de conserver et d'exploiter des corpus polymorphes dont l'état est intrinsèquement non stabilisé, tout en facilitant le versement d'un état stabilisé du corpus vers la BnF. Charge ensuite à la BnF de

40. Cordereix 2005 p. 253-264.

conserver et de donner accès à ce fonds dans le cadre des missions de son service des archives sonores.

Le travail conjoint des chercheurs et des institutions de conservation offre l'opportunité de donner toute sa dimension aux travaux scientifiques fondés sur les données. Il est nécessaire que cette collaboration soit engagée sur l'ensemble d'une chaîne de traitement qui va de la collecte à la conservation en passant par la mise à disposition de ressources libres. Ainsi il convient de ne pas séparer le travail scientifique d'analyse de la gestion des données tant il est vrai que les questions théoriques se posent et s'interpellent à tous les niveaux. C'est le défi que doivent relever tous les acteurs engagés dans de telles recherches.

9. Bibliographie

Baude Olivier et al. Corpus oraux, guide des bonnes pratiques, CNRS et PUO, Paris, 2006.

Paul Cappeau et Magali Seijido, Les corpus oraux en français, DGLFLF, 2005.

CINES, Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles, deuxième édition, 2001.

Cordereix, Pascal, « Les fonds sonores du département de l'audiovisuel de la bibliothèque nationale de France », Le temps des médias n°5. Éditions du nouveau monde, p 253-264, 2005.

DAF/DGME, Standard d'échange de données pour l'archivage, 2006.

Huc Claude, Habert Benoit, Building together digital archives for research in social sciences and humanities, Social Science Information, vol. 49, n°3, p. 415-443, 2010.

IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, Kevin Bradley éditeur. Second edition, 2009.

ISO, Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) - Norme ISO 14721:2003

Jacobson Michel, Corpus oraux en linguistique de terrain, Traitement automatique des langues. 45/2, p. 63-88, 2004.

Jacobson Michel, Que sont les archives ?, lettre d'information de l'InSHS, n°13, 2011.

Olof Bärring, Hosting of IT services and data for Human and Social sciences in France, version 1.0. Rapport final de l'étude commandée par le TGE Adonis au CERN. Document non public, 2008.